Special Review

Artificial Intelligence in Medicine: Where Are We Now?

Sagar Kulkarni, MBBS, BSc, Nuran Seneviratne, MBBS, MA, Mirza Shaheer Baig, MBBS, BSc, Ameer Hamid Ahmed Khan, MBBS, BSc

Artificial intelligence in medicine has made dramatic progress in recent years. However, much of this progress is seemingly scattered, lacking a cohesive structure for the discerning observer. In this article, we will provide an up-to-date review of artificial intelligence in medicine, with a specific focus on its application to radiology, pathology, ophthalmology, and dermatology. We will discuss a range of selected papers that illustrate the potential uses of artificial intelligence in a technologically advanced future.

Key Words: Artificial intelligence; Deep learning; Radiology; Pathology; Ophthalmology; Dermatology; Review.

Abbreviations

AD	adenocarcinoma
AI	artificial intelligence
AUC	area under the receiver-operator curve
CT	computed tomography
DR	diabetic retinopathy
FA	fluorescein angiography
FDA	Food and Drug Administration
ICH	intra-cranial hemorrhage
LYNA	lymph node assistant
MB	megabytes
MRI	magnetic resonance imaging
OCT	optical coherence tomography
OCT-A	optical coherence tomography-angiography
RSNA	Radiological Society of North America
SCC	squamous cell carcinoma
WSI	whole-slide imaging

Acad Radiol 2020; 27:62–70

From the Department of Radiology, Hospital of the University of Pennsylvania, 3400 Spruce Street, Philadelphia 19104, PA (S.K.); Barts and The London School of Medicine and Dentistry, London, United Kingdom (S.K., M.S.B., A.H.A.K.); Department of Geriatric Medicine, The Princess Alexandra Hospital NHS Trust, Harlow, United Kingdom. Received June 23, 2019; revised September 27, 2019; accepted October 1, 2019. **Address correspondence to:** S.K. e-mail: sagar.kulkarni@medsci.ox.ac.uk

https://doi.org/10.1016/j.acra.2019.10.001

INTRODUCTION

Artificial intelligence (AI) is poised to transform medical practice. AI has been studied in several areas of healthcare and medical practice, including precision medicine, population health, and natural language processing (1). The application of AI to visual tasks, known as computer vision, has generated significant interest within the medical community. As such, AI is believed to be relevant to visually-orientated specialties such as radiology, pathology, ophthalmology, and dermatology. The fuel behind AI's development is the availability of large digital datasets; deep learning algorithms use these datasets to train themselves to perform a specific task, such as identifying a lesion in an image. In this article, we review key medical AI studies in the visually-orientated fields with the aim of illuminating the future landscape of AI-enhanced healthcare.

APPLICATIONS IN MEDICAL IMAGING

Plain Film Radiography

The chest radiograph is the most common imaging examination worldwide, with 2 billion performed per year (2). The popularity of chest radiography is explained by its widespread availability around the world and its utility in the diagnosis of a range of conditions. Furthermore, the availability of labeled images, the currency of AI research, is greatest with chest radiographs. For these reasons, chest radiography has garnered the greatest interest amongst AI researchers and continues to be an active research area.

It is fitting to begin the discussion with the data that underpins AI. Among the largest medical AI datasets to date is known as ChestX-ray14. The dataset was released by Wang et al (3) of the National Institutes of Health and consists of 112,120 radiographs from 30,805 unique patients. The images were labeled with 14 conditions, such as emphysema, pulmonary nodules and pneumonia, by four radiologists (three generalists and one thoracic subspecialist). The ground truth was established by a majority vote of the four radiologists. The dataset, originally known as ChestX-ray8, was publicly released in 2017 with the goal of addressing the dearth of labeled data in medical AI research. The dataset is available to use for free, and is still amongst the largest publicly available datasets in the world.

However, ChestX-ray14 has its weaknesses as a dataset, which have been well described in online resources (4,5). For example, diagnostic uncertainty permeates the dataset; practicing radiologists will recognize that there is a level of uncertainty with many radiological diagnoses, and this is evident from the ChestX-ray14 dataset. Wang et al. (3) obtained the ground truth by text-mining through radiology reports. Frequently, these reports contained multiple possible diagnoses, likely because the true diagnosis was radiologically uncertain. Additionally, many of the labels overlap with each other radiologically; for instance, pneumonia can have a similar appearance to atelectasis. Furthermore, there is no definitive evidence affirming whether the radiological diagnosis was correct.

There are further weaknesses with ChestX-ray14, particularly relating to the establishment of the ground truth. Whilst a more detailed discussion of these weaknesses is outside of the scope of this article, we refer the reader to the following resources for further information on this topic (4,5).

ChestX-ray14 was used in a study by Rajpurkar et al. (6) to train an AI detection model called CheXNet. The model was tested on 420 new radiographs, achieving an area under the receiver-operator curve (AUC) of 0.7632 in the first version of the article for the diagnosis of pneumonia, progressing to 0.7680 in the third and most recent version of the model. Furthermore, when considering the full range of 14 diagnoses available in the dataset, CheXNet outperformed previous algorithms that had been derived from the same dataset. Intriguingly, CheXNet's performance mirrored human weaknesses in many respects; the algorithm had much greater accuracy in detecting hiatal hernias, a radiographically distinctive diagnosis, compared to pulmonary infiltration, which is frequently ill-defined.

The latest innovation to ChestX-ray14 and CheXNet is the release of Chester in 2019: a web-delivered disease prediction system (7). The goal of Chester is to deliver the AI model CheXNet, which was trained on ChestX-ray14, to a global userbase. With Chester, the CheXNet code is delivered via a web address to any device connected to the internet, and data processing occurs locally. The idea behind Chester is that a web-delivered system allows for wider distribution of the algorithm, but local processing ensures that patient confidentiality is preserved. Democratizing AI in this fashion may aid in increasing its availability in resource-poor nations.

Moving to AI studies on other datasets, Lakhani and Sundaram investigated the efficacy of a deep learning neural network in detecting features of tuberculosis on chest radiographs (8). After testing multiple deep neural networks, the best-performing classifier had an AUC of 0.99. Additionally, for 13 of the 150 images in the test set where the findings were discordant between the neural networks, the images were referred to a cardiothoracic radiologist, who correctly identified 100% of cases. Therefore, using a collaborative approach between humans and machines delivered a sensitivity of 97.3% and specificity of 100%. The authors suggested that such technology could be used in developing nations with strained resources.

Since 2017, the Radiological Society of North America (RSNA) has led the organization of the Machine Learning Challenge, where research teams from across the world are encouraged to compete to develop the best AI systems for clinical tasks. For the first iteration of the competition, the RSNA challenged researchers to develop deep learning algorithms for pediatric bone age quantification using hand radiographs (9), a common task undertaken by pediatric radiologists. The clinical task was inspired by a paper authored by Larson et al. (10), who developed a deep learning neural network capable of assessing bone age with equivalent accuracy to practicing radiologists. The mean difference in age estimation between the AI network and the human observers was 0 years, with a mean absolute difference of 0.50 years. Ultimately, this algorithm was beaten by all entrants in the competition, with entries bearing a mean absolute difference range of 4.265 to 4.907 months on the same dataset of 12,611 hand radiographs; the top three performers were separated by 3.5 days (11).

In response to the growing interest in AI applications in radiology, the RSNA established *Radiology: Artificial Intelligence*. In the journal's first issue, Thian et al. (12) made the front cover for their work on fracture detection. The authors used ResNet, a pretrained deep learning model, to devise an AI system for ulnar and radial wrist fracture detection. The authors used 7,356 wrist radiograph studies, of which 90% were used for training and 10% used for validation, to develop the model. The images were annotated by radiologists placing bounding boxes around suspected fractures. When the model was trialed on 524 consecutive emergency department wrist radiographs (with two radiologists as a reference standard), the model correctly localized 91.2% of radial fractures and 96.3% of ulnar fractures. Per radiographic study, the sensitivity, specificity and AUC were 98.1%, 72.9%, and 0.895, respectively. Since wrist fractures are amongst the most common fractures encountered in the emergency department (13), the wrist fracture model may lend itself well to automatic detection and triaging in the future.

Advanced Imaging

AI also has several applications in advanced imaging. For instance, magnetic resonance imaging (MRI) may be virtually enhanced using AI models. Gong et al. (14) trained a deep learning algorithm on brain MR images on 10 pre- and postcontrast brain MR images, allowing it to learn how the image changes after the administration of gadolinium. The algorithm was then applied to a series of low-dose contrast-enhanced images, where it virtually enhanced the contrast of the gadolinium present in the image. The use of the algorithm provided significant improvements in peak-signal-to-noise ratios of more than 5 decibels, allowing greater interpretability of the images. Furthermore, a noninferiority analysis revealed that image quality, artifact suppression, and contrast enhancement was not significantly different between full-dose and low-dose contrast images. In the wider medical community, controversy exists over the administration of gadolinium due to concerns of its effects on the renal system and the brain; (15) applying deep learning to this problem could reduce the required dose of gadolinium tenfold (14).

AI has also been used under experimental conditions as an end-to-end reader of screening examinations. Ardila et al. (16), a team composed of investigators from Google AI (Mountain View, CA) and multiple US hospitals, used a dataset of 42,290 publicly available lung cancer screening computed tomography (CT) scans from the National Lung Cancer Screening Trial to train, tune and test a deep learning architecture. On testing with 6,716 scans, the algorithm achieved an AUC of 0.944, outperforming 6 board-certified radiologists when no prior imaging was available, and equaling their performance with prior imaging. This technology could be utilized in a public health setting; currently, although lung cancer is the most common cause of cancer death in the US (17), screening has a relatively poor uptake. AI could reduce the cost of screening, thereby encouraging participation in early detection programs.

As previously noted, a common limitation of AI research is a lack of appropriately labeled data; however, AI itself may be the solution to this issue. Using deep learning in brain MRI, Dalca et al. (18) demonstrated an algorithm that combined deep learning with a traditional probabilistic atlas to drive a brain segmentation algorithm. Using a training dataset of 7,332 brain MR scans, the newly developed deep learning algorithm outperformed the baseline Gaussian likelihood functions overall (Dice score 83.5% vs 79.0%, where a high Dice score indicates higher similarity between the segmentation method and the ground truth), and was particularly successful at segmenting deep brain structures such as the hippocampus (Dice score 81.1% vs 73.1%). The value of such a technique is that it could remove the time-consuming, resource-intensive process of labeling images, thereby increasing the numerical size of datasets that can be used for AI research.

Noninterpretive Tasks

Perhaps AI's greatest utility will be outside of image interpretation entirely. A radiology fellow spends only 53.8% of their working time on image interpretation (19). The remainder of the time is spent on nonimage-interpretative tasks, such as protocoling studies, consulting with technologists, and consulting with clinicians regarding critical findings (19). AI could assist in improving workflow by aiding radiologists in these tasks.

One of the first steps in the imaging pathway involves protocoling the study. Important considerations to be made are; whether the study is clinically indicated, determining whether the appropriate acquisition parameters (for example, the pulse sequence) are correct and whether contrast administration is required. Radiologists spend 6.2% of their time protocoling studies (19). A deep learning algorithm has been developed to improve this process; Lee (20) used a dataset of 5,258 musculoskeletal MR requests to train a deep learning network to appreciate factors such as the word combination of the request, the use of contrast media and the demographic characteristics of the patient. The study reported that the algorithm had an AUC of 0.977 in determining the correct protocol for imaging. Implementing such an algorithm in a busy radiology department could improve productivity through workflow efficiency savings.

Furthermore, AI could also improve the acquisition of scans. One of the major weaknesses of MRI as a modality is its lengthy acquisition time. The prolonged acquisition time is uncomfortable for patients and makes motion artifacts more likely to occur. Furthermore, longer acquisition times reduce the throughput of the scanner. Hyun et al. (21) developed an algorithm capable of reconstructing undersampled MRI data to form a full quality image. The algorithm used a training set of 1,400 sets of undersampled and fully sampled images. The authors showed both qualitatively and quantitatively that only 29% of k-space data was required to reconstruct a full quality

MR image. Since much of the reduction in k-space acquisition was made by undersampling during acquisition in the phase encoding direction, an MR scan could be completed much more rapidly using this technique.

Following acquisition, AI has also been used to assist in worklist prioritization. Arbabshirani et al. (22) used a dataset of 37,074 head CT scans to train an algorithm to detect the presence of intracranial hemorrhage (ICH). If an ICH was detected, the model would prospectively update the priority of the scan to "stat." The algorithm achieved an AUC of 0.846 for the detection of ICH. Once the algorithm was prospectively implemented into real-world clinical workflow, 27% of routine of inpatient and outpatient (but excluding emergency department) head CT requests were reprioritized to "stat." Of these reprioritized scans, 85% were diagnosed with ICH by a radiologist, equating to a reduction in the time to diagnosis from 512 minutes to 19 minutes between routine and stat studies. Survival from ICH is dependent on rapid identification of the condition, therefore, using an algorithm to prioritize an emergency worklist may have beneficial outcomes for patients.

Years after the examination, AI may still have utility to harness information from the report. An AI algorithm has been developed to classify free-text radiology reports. The model, developed by Chen et al. (23), classifies the presence, location and chronicity of pulmonary emboli on thoracic CT scans based solely on the text of the report. After training on 2,500 reports, the model achieved an AUC of 0.97. The retrospective classification of reports could pave the way for more analytic research in radiology and in medicine overall, which is currently restricted by presently used electronic health records systems. Furthermore, the study reveals an alternative by which further labeled data could be obtained for future AI studies.

APPLICATIONS IN PATHOLOGY

Unlike the progression of radiology from illuminated X-ray films to digital imaging, pathology has progressed at a slower pace to the digital medium, which the adoption of AI hinges on. Whole-slide imaging (WSI) now enables pathologists to view histopathology slides in their entirety in high resolution with depth manipulation. Despite the availability of WSI and its benefits, digital conversion of glass slides is not routinely carried out (24). Rapid advances in technology have enabled fast transfers and ample storage for vast amounts of data. This is particularly enabling in the field of WSI where the amount of data is large and requires real time processing. A typical size for a whole-slide scan can be around 1.6 billion pixels taking up around 4600 MB (megabytes) of storage space (25). On the other hand, radiology images shared in DICOM (Digital Imaging and Communications in Medicine) format range from approximately 4000 pixels for some nuclear medicine scans to 23 million pixels in mammography images (26). The storage size of these images depends on the number of images per study but MRIs can typically take around 30−50 MB of storage space as demonstrated on the online DICOM library (27). File size and resolution difference between radiology and pathology studies could be an important differentiating factor in the adoption of AI in the future.

The question to be answered is how proficient artificial intelligence is in providing vital analytical ability in digital pathology. Although available literature is limited, there are a handful of studies comparing the performance of AI algorithms with pathologists in detection and classification of different types of cancer.

In a retrospective study of breast cancer metastases, Ehteshami Bejnordi et al. (28) investigated the performance of 32 submitted deep learning algorithms in detecting lymph node metastases in WSI of tissue sections from women with breast cancer compared to a panel of expert pathologists. An annotated training data set was provided to develop the challenge participants' algorithms which included 110 images with nodal metastases and 160 without. The performance test included 129 images of which 49 were verified with metastases and 80 without. The assessment was split into two tasks: The first task to identify individual metastases and the second to classify metastases. The same images were presented to a panel of 11 pathologists under a time constraint as well a single pathologist without time constraints. The pathologists with time constraints were given a flexible time limit of 2 hours to review the 129 slides while the pathologist without time constraints used 30 hours.

The top scoring algorithm performed significantly better in the image classification task with an AUC of 0.994 compared to the pathologists with time constraints, mean AUC of 0.810. However, it was comparable to the single pathologist without time constraints with an AUC of 0.966. These results show that, while a pathologist utilizing 30 hours to assess 129 slides is clinically impractical, this level of performance can be achieved with a deep learning algorithm. However, it must be noted that pathologists with time constraints in this study had less than 1 minute per slide (although flexible) which may be a poor reflection of true clinical practice. In addition, the same pathologists used glass slides rather than digital WSI which may yield more information albeit in an unfamiliar format. Nevertheless, this study highlights the potential for AI to provide fast and reliable analysis of tissue samples to detect breast cancer metastases at a performance level of a seasoned pathologist with an unlimited amount of time.

Furthermore, Coudray et al. (29) investigated the performance of deep learning algorithms in classification and mutation prediction in non-small cell lung cancer histopathology images. In this study, deep learning algorithms and pathologists were asked to classify and distinguish between adenocarcinoma (AD) and squamous cell carcinoma (SCC) in tissue sample slides. The team trained a deep learning algorithm using WSIs obtained from the cancer genome atlas to reliably differentiate between AD, SCC and normal tissue. The dataset used in this study consisted of 459 normal tissue slides, 567 classified as AD and 608 as SCC. These annotations were provided by the cancer genome atlas. These slides were split between training, validation and testing sets. In comparison to the study by Ehteshami Bejnordi et al (28), the WSIs used here were deemed to be too large to be inputted directly into the deep learning algorithm and they used 512 × 512 pixel tiles instead.

The results of this study indicated comparable performance between AI and pathologists. Although the AUC of the deep learning model was higher in identification and classification of lung cancer samples (0.99 and 0.97 AUC respectively) compared to the performance of three pathologists, the comparison did not reach the threshold of statistical significance to conclude a superior diagnostic performance. However, this study did highlight the ability of this deep learning algorithm to predict gene mutational status from WSIs, particularly STK11 mutations which were predicted with the highest accuracy (AUC 0.85) that would not otherwise be detected by pathologists – showing that AI can be a powerful tool in supplementing pathologists' diagnoses.

Classification of brain tumors is a difficult task, with studies reporting significant inter-observer variability in histopathological diagnosis of various CNS tumors (30). This is largely a result of the diverse nature of these tumors. The current World Health Organization classification of CNS tumors includes a vast number of different tumors which arise from the developmental complexity of the brain (31). Capper et al. (32) used tumor DNA methylation as a distinguishing measure to classify different CNS tumors. DNA methylation profiles generate vast amounts of information which are not routinely used in clinical practice but this study highlighted the processing power of machine learning in classifying CNS tumors using DNA methylation profiles. From a prospective analysis of 1155 samples, 838 (76%) were successfully classified by the machine learning program using DNA methylation profiles with 129 (12%) samples being classified as establishing a new diagnosis with a large clinical impact. Although not a direct comparison of performance with current pathologists, this study points to the potential for AI to provide powerful, precise and wide-reaching diagnostic ability in the future.

Steiner et al. (33) investigated the impact of AI assistance in histopathological review of lymph node biopsies for metastatic breast cancer. In comparison to the study by Ehteshami Bejnordi et al (28), they compared pathologist performance assisted versus non-assisted by AI to determine the potential benefit from supplementation. They developed the deep learning algorithm, Lymph Node Assistant (LYNA) from the same dataset used in the study by Ehteshami Bejnordi et al. (28) and described in Liu et al. (34) Pathologists were randomly assigned into a cohort with crossover between each cohort in receiving LYNA assistance or unassisted. A total of 70 digitized images were formally reviewed as part of the study. The study showed AI assistance increased both accuracy, sensitivity, and time efficiency when compared to the unassisted cohort. Operating points for the sensitivity and specificity were higher for all algorithm-assisted pathologists than the receiver-operator curve generated for LYNA alone; however, some unassisted pathologists were less sensitive than LYNA alone. Overall, algorithm-assisted pathologists performed particularly well in the detection of micrometastases compared to unassisted (AUC 91% versus 83%) and were more time efficient in detecting these findings (61 vs 116 seconds).

Steiner et al. highlight the importance in considering AI as a synergistic tool in clinical practice in improving patient outcomes. They show that algorithm-assistance has the potential to not only provide more accurate and sensitive histopathological diagnoses but also save time and increase clinician confidence, all of which contribute to improved patient outcomes.

APPLICATIONS IN OPHTHALMOLOGY

Diabetic eye disease is amongst the most common conditions seen in routine ophthalmology practice, and constitutes a significant and growing public health issue. Diabetic retinopathy (DR) is the commonest cause of vision loss in working age adults, with 2.6 million people affected globally in 2015, expected to rise to 3.2 million in 2020 (35). The incidence of sight threatening diabetic retinopathy in upper income countries is falling, by a combination of both better diabetic control and ophthalmological interventions, however this is offset worldwide by increased diabetes incidence and increasing incidences of DR in lower resources countries (36). With outcomes improving dramatically with early detection and with the widening provision of digital imaging (high-resolution color retinal photography and optical coherence tomography), coupled with the potential for irreversible morbidity, DR is an ideal candidate for to be screened for. However, with strict guidelines and a limited workforce to effect screening has led to a demand for streamlining pathways to a specialist ophthalmologist review.

It is that demand that AI diagnostics company IDx (Coralville, IA) aims to fulfill, gaining Food and Drug Administration (FDA) approval its IDx-DR cloud-based AI system in April 2018. It is the first medical device to be authorized to provide a screening decision without the oversight of a clinician, stratifying patients into those that have "more than mild" DR who require ophthalmology review, and those that do not, who require 12 monthly screening.

Critical to this decision was the first prospective trial of the system in 10 primary care settings by Abramoff et al. of 892 patients yielding 819 analyzable images, of which IDx-DR achieved 87.2% sensitivity, and 90.7% specificity in its classification task (37). These clinical successes build upon the system achieving an AUC of 0.94 in discriminating referable diabetic retinopathy in a 1415 patient retrospective dataset, the Hoorn Diabetes Care System Cohort, of which 898 were analyzable (38). Other retrospective studies show sensitivities of over 91% (39,40), important metrics in this screening investigation where false negatives have potential for harm but false positive findings merely lead to inappropriate review by an ophthalmologist.

However, the prospective Abramoff trial is not without fault, as a low incidence of more than mild DR in study populations required pre-enrichment strategies to recruit higher glycated hemoglobin (A1C) and fasting blood sugar patients. This may be inevitable when designing prospective studies on AI systems in screening programs, which by their nature must trawl through large asymptomatic populations to capture a relatively small population in the detectable presymptomatic phase.

Lack of human expert consensus reflects the difficulties in DR diagnosis, and mechanisms must exist to ameliorate these disagreements in order to produce an accurate reference standard upon which these algorithms are trained. Ground truth is established predominantly by majority decision of multiple independent experts, or discrepancies arbitrated by a further expert whose decision acts as the reference DR grade. Krause et al. challenged this paradigm in order to improve upon a deep learning algorithm presented by Google to detect referable vs. nonreferable DR, which, in this binary task, had already achieved AUC of 0.991 in its larger data set (EyePACS-1) and AUC of 0.990 in the smaller dataset (Messidor-2) (41). Krause et al. uses a face-to-face adjudication system where graders discuss the discrepancy in grade until a consensus grading is achieved. The authors moved the algorithm from the binary referability task to the widely used 5-point International Clinical Diabetic Retinopathy scale, from no DR to proliferative DR. To adjudicate the more than 1.6 million training images would have been unfeasible, and so a 3737 image adjudicated tuning set was used to adjust the algorithms hyper-parameters as well as shape modeling choices. Their algorithm had high agreement to the reference standard, with similar kappa values to the human experts (retinal specialists 0.82 to 0.91, general ophthalmologists 0.80 to 0.84, and the algorithm 0.84) (42). However, other algorithmic improvements may contribute to this, such as using higher resolution input images and changing to a more contemporary model architecture.

The utility of this algorithm to supplement clinician grading is explored in a study measuring the time, accuracy and confidence of DR grading by 10 ophthalmologists in 3 levels of algorithm assistance; no assistance, when provided with the algorithm-derived DR grade, and when provided with this grade and a heat map of areas that contributed to the algorithms grading. Interestingly, accuracy increased across the board with the grades alone ($P < 0.001$), yet no increase in accuracy was seen when grades and the heat-map were provided. Human grading time increased predominantly early in the study suggesting that with experience grading time decrease. This raises questions as to how additional algorithm-derived information is perceived by human graders, and the authors theorise that the heat maps may cause second-guessing of normal imaging, as there was a significant decrease in accuracy when heat maps were provided in images without DR ($P = 0.007$), leading to overcalling of DR. Grading of no DR is usually accurate, ranging from 92.5−94.7% accuracy across all readers and algorithm assistance levels. In addition, there was a marked increase in reporting feeling extremely confident in their grading when the algorithm grade was provided, whereas with the heat map the increased confidence was spread between very and extremely confident (43). This paper exemplifies how algorithm derived information can both improve the accuracy of human diagnosis, but may inadvertently confound otherwise accurate decision making processes.

AI algorithms have also been developed for optical coherence tomography (OCT), another diagnostic technique in ophthalmology. One such framework (44) uses a two-step process to identify pathology. First, via a deep segmentation network, image acquisition related variations are identified; then, a deep classification network is applied to the resultant segmentation map to identify pathology. Of 53 key diagnoses found in total, the most urgent diagnosis identified was used to form a referral recommendation based on the Moorfields Eye Hospital (London, United Kingdom) referral criteria-urgent, semi-urgent, routine or observation only. When compared to 8 clinical experts (4 retinal subspecialist ophthalmologists and 4 optometrists with additional medical retina training), it achieved an AUC of 0.9921 in identifying urgent referral cases, matching the performance of 2 ophthalmologists, and outperforming all other experts. When the experts were given access to supplementary information (fundus images and clinical notes) this number rose to 5 experts that the system demonstrates noninferiority to. Total referral misclassification rates were low at 5.5%, matching 2 experts and surpassing the others.

The authors attribute this accuracy to the use of multiple neural networks, 5 each for segmentation and classification, analogous to a panel of experts. As expected, the authors found multiple neural networks found to be superior to using a single network. Extraordinarily, areas of ambiguity where the networks disagree, and propose multiple hypotheses, can be illustrated as a video. This can then be used to guide clinical decision-making by the ophthalmologist, and reflects the lack of consensus in human experts (all 8 experts agreed on only 65% of images).

In addition to screening and diagnostic tasks, AI has been used in ophthalmology to infer additional clinically significant information. Optical Coherence Tomography Angiography (OCT-A) is emerging as a non-invasive alternative to fluorescein angiography (FA) in mapping the retinal vasculature by acting as a motion-contrast detector between dynamic blood and static neurosensory tissue (45). However, its widespread usage is limited by cost and requirement for patient collaboration in this extremely motion sensitive modality, as well as limited fields of view (46). After training a deep learning model on 400,000 OCT images and their corresponding OCT-A, the model was able to detect microvasculature, large and medium vessels from OCT at a similar level of detail to the corresponding OCT-A. Using OCT-A as ground truth, when the model was compared to three clinicians segmenting vasculature from OCT, the AI significantly outperformed the experts (47). As clinicians would refer for OCT-A or FA for the definitive diagnosis of micro-vascular complications based on their assessment of the OCT, this comparison is not strictly a fair one if putting forward the AI as an alternative to OCT-A or FA. However, the model may be useful in referring to these definitive diagnostic tests, and if put forward as an alternative to the definitive tests, it could ameliorate drawbacks of not just contrast dye injection but technical limitations of OCT-A. This study exemplifies the potential of AI to accurately infer useful functional data from structural data.

Fundamentally, it has yet to be ascertained whether these new technologies will improve provision and uptake of diabetic retinal screening, diagnosis and appropriate referral of other sight threatening disorders, or replace current investigations. Yet these technologies may go on to form the basis of paradigm shifts in future ophthalmological practice.

APPLICATIONS IN DERMATOLOGY

The recognition of visual patterns is a fundamental diagnostic skill in dermatology and AI may provide much potential in augmenting image analysis and improving diagnostic accuracy within this field (2,48). Recently developed computational neural networks have been used to diagnose skin conditions through visual image recognition and have demonstrated comparable and occasionally greater sensitivity and specificity in classifying images than even clinically experienced dermatologists.

For instance, Esteva et al. used the GoogLeNet Inception algorithm, which was pretrained on 1.28 million images and then retrained on a university dataset of 129,450 high quality dermatological images (49). The patterns of submitted digital images were then analyzed at a pixel level and given a diagnosis. Twenty-one US board-certified dermatologists were matched or exceeded by the deep learning algorithm, which had an AUC of 0.96 for carcinoma and 0.94 specifically for melanoma. Haenssle et al. compared the diagnostic accuracy of a convolutional neuronal network with an international group of 58 dermatologists, which included 30 experts, and found that most of the dermatologists were outperformed. The convolutional neuronal network group had a higher area under the ROC curve of 0.86 compared with 0.79 in the dermatologists group (50). In another study, Brinkler et al. had shown that convolutional neuronal networks were able to outperform 136 of 157 dermatologists in classifying 12,378 head-to-head dermatoscopic images of suspicious skin lesions including melanoma (51).

Convolutional neuronal networks have also been used to classify clinical images of skin diseases beyond skin cancers. Han et al. conducted an algorithmic assessment of 12 skin diseases which included actinic keratoses, seborrheic keratoses, melanocytic nevi, pyogenic granulomas, hemangiomas, and warts in addition to common skin cancers (52). The convolutional neuronal network, Microsoft ResNet-152 model (Microsoft Research Asia, Beijing, China) was trained with 19,838 images from the training segment of the Asan dataset, MED-NODE dataset and atlas site images. This trained model was then validated with the testing segment in 3 datasets including the Asan and Edinburgh datasets. For images in the Asan set, the AUC curve was 0.82−0.95, the sensitivity was 77.7−93.9% and the specificity was 74.3−92.6%. For images in the Edinburgh set, the area under the curve was 0.83−0.97, sensitivity was 77.3−98.6% and specificity was 70.5−89.6%. The study showed that the algorithm's performance using 480 Asan and Edinburgh images were similar to the performance of 16 compared dermatologists.

AI may also play a future role in the earlier detection and treatment of skins cancers. Chuchu et al. examined evidence for the potential use of smartphone technology in providing lay users an early risk assessment tool for melanoma, which accounts for most skin-cancer related deaths despite only forming a small proportion of all skin cancers (53). Across the 4 AI-based applications that they examined, sensitivities ranged from 7−73% and specificities and specificities 37−94%. The number of skin lesions classed as unevaluable ranged from 2% to 18% of all lesions analyzed and in 3 of the 4 applications at least one melanoma was classified as unevaluable. Thus, to date, AI has not demonstrated sufficient diagnostic accuracy in smartphone technology and given the high likelihood of missing melanomas, may not be yet suitable as means of self-screening in a lay population. However, since there are a lack of studies and there is a rapid rate of development within this field, further studies in the future may yield more promising results in combining smartphone technology with AI to detect melanomas (54).

The use of AI in dermatology also poses significant challenges. Whilst AI may also have the potential to improve workflow efficiency, difficulties have also been identified in integrating AI into conventional clinical workflow systems which, at least in the foreseeable future, is likely to impede its clinical use for improving diagnostics in dermatology (55).

Within imaging analysis, dermatology remains a particularly challenging field for the application of AI. Unlike radiological imaging, there is a lack of standardization in skin imaging with regards to color, lighting, techniques and hardware (55). Skin tone can also have a considerable effect on the image appearance.

Furthermore, vast sets of annotated, high-resolution imaging data, encompassing the breadth of diagnostic variety, are required to establish a ground truth for the automated development of algorithms in deep neural networks (55). Compiling these data sets can be both time intensive and expensive. Nevertheless, as in other applied contexts, AI itself may be the answer to saving time and expense in establishing the ground truth. Zhang et al. demonstrated in their study that their machine learning model, known as multi-instance multilabel, was able to annotate skin biopsy images using an algorithm and was shown to be effective on a clinical data set consisting of 12,700 biopsy images (56). Accurate and rapid annotation of skin imaging could thus be used to establish a ground truth for computational neuronal networks in classifying new clinical images.

Concerns have also been raised by dermatologists regarding the use of AI within the specialty. Lim and Flaherty acknowledge that while AI may improve the diagnostic accuracy and efficiency, it warns of any blind adoption of the technology where other holistic diagnostic aids such as history and clinical context are ignored (57).

CHALLENGES

There will be several challenges to the implementation of AI in healthcare. For illustrative purposes, we will focus on three that relate to the studies discussed previously in this paper: The black box problem, overfitting, and regulatory approval.

The black box problem is the inability of deep learning algorithms to demonstrate how they arrive at their conclusions. When an algorithm infers a radiological finding, traditionally it has been impossible to determine which imaging features were used in the process, how these were analyzed and why the algorithm arrived at one outcome over another. Most of the AI systems discussed in this paper exhibit the same problem. In select cases, attempts have been made to overcome this issue. For instance, Lakhani and Sundaram (8), in their paper on deep learning in tuberculosis, used a heat map to show regions of increased activation of the deep learning network, which we may infer are regions of high importance in determining the diagnosis. Such methods go some way toward opening the black box of AI.

Overfitting is when AI algorithms, trained on one dataset, have limited applicability to other datasets. This is because the algorithm has learned the statistical variation of the training data, as opposed to the broad concepts needed to solve a problem. The key determinant of overfitting is the overtraining of an algorithm on a specific dataset. Several factors influence the likelihood of overfitting, including the size of the dataset, the extent of heterogeneity within the dataset and the distribution of the data within the dataset. For instance, a model may be overfitted if the prevalence and incidence of disease differs significantly between the training and testing sets, or if the training and testing sets were acquired with substantially different parameters or equipment, which may be further compounded by small sample size. Following training, algorithms may be examined for overfitting by testing them on multiple different datasets; in an algorithm suffering from overfitting, one would expect its accuracy, measured by the AUC, to be significantly worse on datasets that do not bear the same origin as the training data (58).

Finally, regulatory approval will pose a challenge for new AI algorithms. Medical AI, like drugs and medical devices, will be regulated by the FDA. Both the black box problem and overfitting combine to create barriers for regulatory approval, since evaluators face difficulties in determining how the algorithms work and whether their performance is generalizable to other datasets. The FDA classifies new AI tools on the basis of three criteria: risk to patient safety, the existence of a predicate algorithm and the degree of human input (59). Algorithms that are deemed to be high risk, such as diagnostic tools where the consequences of a misdiagnosis are severe and where there is minimal human input, undergo evaluation by the premarket approval pathway, which requires substantial evidence from non-clinical and clinical studies that the new tool is safe and efficacious. Some lower risk technologies may be evaluated by the De Novo pathway, which is designed for the approval of revolutionary devices. For a more extensive discussion of FDA regulation and approval of new AI tools in medical imaging, we refer the reader to the following review by Kohli et al. (59).

CONCLUSION

AI has several potential applications in medicine; it remains to be seen which of these will take hold. Certainty, however, lies in the inevitability of change. Therefore, it is important for all physicians to be aware of the recent advances of AI, as it is likely to influence the delivery of healthcare in the future.

REFERENCES

1. Davenport T, Kalakota R. The potential for artificial intelligence in healthcare. Futur Healthc J 2019; 6(2):94. [cited 2019 Sep 23]. Available from: http://www.ncbi.nlm.nih.gov/pubmed/31363513.
2. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med 2019. [cited 2019]. Available from: doi:10.1038/s41591-018-0300-7.
3. 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. 2017. [cited 2019 May 9]. Available from: http://arxiv.org/abs/1705.02315
4. Oakden-Rayner L. Exploring the ChestXray14 dataset: problems. 2017 [cited 2019 Sep 23]. Available from: https://lukeoakdenrayner.wordpress.com/2017/12/18/the-chestxray14-dataset-problems/
5. Borstelmann S.CheXNet - a brief evaluation. Volume to Value. 2017. [cited 2019 Sep 23]. Available from: https://n2value.com/blog/chexnet-a-brief-evaluation/
6. Rajpurkar P, Irvin J, Zhu K, et al. CheXNet: radiologist-level pneumonia detection on chest x-rays with deep learning. 2017; Available from: http://arxiv.org/abs/1711.05225
7. Cohen JP, Bertin P, Frappier V. Chester: a web delivered locally computed chest x-ray disease prediction system. 2019; Available from: http://arxiv.org/abs/1901.11210
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. [cited 2019 May 9]. Available from: http://pubs.rsna.org/doi/10.1148/radiol.2017162326.
9. Halabi SS, Prevedello LM, Kalpathy-Cramer J, et al. The RSNA pediatric bone age machine learning challenge. Radiology 2019; 290(2):498–503. Available from: http://pubs.rsna.org/doi/10.1148/radiol.2018180736.
10. Larson DB, Chen MC, Lungren MP, et al. Performance of a deep-learning neural network model in assessing skeletal maturity on pediatric hand radiographs. Radiology 2018; 287(1):313–322. Available from: http://www.ncbi.nlm.nih.gov/pubmed/29095675.
11. Siegel EL. What can we learn from the RSNA pediatric bone age machine learning challenge? Radiology 2019; 290(2):504–505. Available from: http://pubs.rsna.org/doi/10.1148/radiol.2018182657.
12. Thian YL, Li Y, Jagmohan P, et al. Convolutional neural networks for automated fracture detection and localization on wrist radiographs. Radiol Artif Intell 2019; 1(1):e180001. [cited 2019 Sep 23]Available from: http://pubs.rsna.org/doi/10.1148/ryai.2019180001.
13. Voth M, Lustenberger T, Auner B, et al. What injuries should we expect in the emergency room? Injury 2017; 48(10):2119–2124. [cited 2019 Sep 24]. Available from: http://www.ncbi.nlm.nih.gov/pubmed/28778731.
14. Gong E, Pauly JM, Wintermark M, et al. Deep learning enables reduced gadolinium dose for contrast-enhanced brain MRI. J Magn Reson Imaging 2018; 48(2):330–340.
15. Guo BJ, Yang ZL, Zhang LJ. Gadolinium Deposition in Brain: Current Scientific Evidence and Future Perspectives. Front Mol Neurosci 2018; 11:335. [cited 2019 May 13]. Available from: http://www.ncbi.nlm.nih.gov/pubmed/30294259.
16. 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. Available from: http://www.nature.com/articles/s41591-019-0447-x.
17. American Lung Association. Lung cancer fact sheet. American Lung Association.

18. Dalca A V, Yu E, Golland P, et al. Unsupervised deep learning for Bayesian brain MRI segmentation. J Magn Reson Imaging 2019; 48(2):330–340. Available from: http://arxiv.org/abs/1904.11319.
19. Schemmel A, Lee M, Hanley T, et al. Radiology workflow disruptors: a detailed analysis. J Am Coll Radiol 2016; 13(10):1210–1214.
20. Lee YH. Efficiency improvement in a busy radiology practice: determination of musculoskeletal magnetic resonance imaging protocol using deep-learning convolutional neural networks. J Digit Imaging 2018; 31 (5):604–610.
21. Hyun CM, Kim HP, Lee SM, Lee S, Seo JK. Deep learning for undersampled MRI reconstruction. Phys Med Biol 2018; 63(13):135007. Available from: http://www.ncbi.nlm.nih.gov/pubmed/29787383.
22. Arbabshirani MR, Fornwalt BK, Mongelluzzo GJ, et al. Advanced machine learning in action: identification of intracranial hemorrhage on computed tomography scans of the head with clinical workflow integration. NPJ Digit Med 2018; 1(1). Available from: http://dx.doi.org/10.1038/s41746-017-0015-z.
23. Matthew C, Chen M, Robyn L, et al. Deep learning to classify radiology free-text reports. Radiology. 2017; 286(3):845–852.
24. Acs B, Rimm DL. Not just digital pathology, intelligent digital pathology. JAMA Oncol 2018; 4(3):403. [cited 2019 May 23]. Available from: http://www.ncbi.nlm.nih.gov/pubmed/29392271.
25. Pantanowitz L, Farahani N, Parwani A. Whole slide imaging in pathology: advantages, limitations, and emerging perspectives. Pathol Lab Med Int 2015; 7:23. [cited 2019 Sep 23]. Available from: http://www.dovepress.com/whole-slide-imaging-in-pathology-advantages-limitations-and-emerging-p-peer-reviewed-article-PLMI.
26. Kohli MD, Summers RM, Geis JR. Medical image data and datasets in the era of machine learning—whitepaper from the 2016 C-MIMI meeting dataset session. J Digit Imaging 2017; 30(4):392–399. [cited 2019 Sep 23]. Available from: http://link.springer.com/10.1007/s10278-017-9976-3.
27. DICOM Library [cited 2019 Sep 23]. Available from: https://www.dicomlibrary.com/about/
28. Ehteshami Bejnordi B, Veta M, Johannes van Diest P, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. JAMA 2017; 318(22):2199. [cited 2019 May 23]. Available from: http://jama.jamanetwork.com/article.aspx?doi=10.1001/jama.2017.14585.
29. Coudray N, Ocampo PS, Sakellaropoulos T, et al. Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. Nat Med 2018; 24(10):1559–1567. [cited 2019 May 23]. Available from: http://www.ncbi.nlm.nih.gov/pubmed/30224757.
30. van den Bent MJ. Interobserver variation of the histopathological diagnosis in clinical trials on glioma: a clinician's perspective. Acta Neuropathol 2010; 120(3):297–304. [cited 2019 May 23]. Available from: http://link.springer.com/10.1007/s00401-010-0725-7.
31. Louis DN, Perry A, Reifenberger G, et al. The 2016 World Health Organization Classification of tumors of the central nervous system: a summary. Acta Neuropathol 2016; 131(6):803–820. [cited 2019 May 23]. Available from: http://link.springer.com/10.1007/s00401-016-1545-1.
32. Capper D, Jones DTW, Sill M, et al. DNA methylation-based classification of central nervous system tumours. Nature 2018; 555(7697):469–474. [cited 2019 May 23]. Available from: http://www.nature.com/articles/nature26000.
33. Steiner DF, MacDonald R, Liu Y, et al. Impact of deep learning assistance on the histopathologic review of lymph nodes for metastatic breast cancer. Am J Surg Pathol 2018; 42(12):1636–1646. [cited 2019 May 23] Available from: http://www.ncbi.nlm.nih.gov/pubmed/30312179.
34. Liu Y, Gadepalli K, Norouzi M, et al. Detecting cancer metastases on gigapixel pathology images. 2017[cited 2019 May 23]; Available from: http://arxiv.org/abs/1703.02442
35. Flaxman SR, Bourne RRA, Resnikoff S, et al. Articles global causes of blindness and distance vision impairment 1990-2020: a systematic review and meta-analysis. 2017[cited 2019 May 23]; Available from: www.thelancet.com/lancetgh
36. Cheloni R, Gandolfi SA, Signorelli C, et al. Global prevalence of diabetic retinopathy: protocol for a systematic review and meta-analysis. BMJ Open 2019; 9(3):2015–2019.
37. Abràmoff MD, Lavin PT, Birch M, et al. Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. NPJ Digit Med. 2018; 1(1):39.
38. van der Heijden AA, Abramoff MD, Verbraak F, et al. Validation of automated screening for referable diabetic retinopathy with the IDx-DR device in the Hoorn Diabetes Care System. Acta Ophthalmol. 2018; 96 (1):63–68.
39. Hansen MB, Abràmoff MD, Folk JC, et al. Results of automated retinal image analysis for detection of diabetic retinopathy from the Nakuru Study, Kenya. PLoS One 2015; 10(10):e0139148.
40. Grinsven M. Van. Automated analysis of retinal images for detection of age-related macular degeneration and diabetic retinopathy. 2017.
41. 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.
42. Krause J, Gulshan V, Rahimy E, et al. Grader variability and the importance of reference standards for evaluating machine learning models for diabetic retinopathy. Ophthalmology 2018; 125(8):1264–1272.
43. Sayres R, Taly A, Rahimy E, et al. Using a deep learning algorithm and integrated gradients explanation to assist grading for diabetic retinopathy. Ophthalmology 2019; 126(4):552–564.
44. De Fauw J, Ledsam JR, Romera-Paredes B, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. Nat Med 2018; 24(9):1342–1350.
45. Kashani AH, Chen CL, Gahm JK, et al. Optical coherence tomography angiography: a comprehensive review of current methods and clinical applications. Prog Retin Eye Res 2017; 60:66–100.
46. De Oliveira PRC, Berger AR, Chow DR. Optical coherence tomography angiography in chorioretinal disorders. Can J Ophthalmol 2017; 52 (1):125–136.
47. Lee CS, Tyring AJ, Wu Y, et al. Generating retinal flow maps from structural optical coherence tomography with artificial intelligence. Sci Rep 2019; 9(1):1–11.
48. Schlessinger DI, Chhor G, Gevaert O, et al. Artificial intelligence and dermatology: opportunities, challenges, and future directions. Semin Cutan Med Surg 2019; 38(1):E31–E37.
49. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017; 542(7639):115–118. [cited 2019 May 8]. Available from: http://www.nature.com/articles/nature21056.
50. Haenssle HA, Fink C, Schneiderbauer R, et al. Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. Ann Oncol 2018; 29(8):1836–1842. [cited 2019 May 12]. Available from: https://academic.oup.com/annonc/article/29/8/1836/5004443.
51. Brinker TJ, Hekler A, Enk AH, et al. Deep learning outperformed 136 of 157 dermatologists in a head-to-head dermoscopic melanoma image classification task. Eur J Cancer 2019; 113:47–54.
52. Seog Han S, Shin Kim M, Lim W, et al. Classification of the clinical images for benign and malignant cutaneous tumors using a deep learning algorithm. J Invest Dermatol 2018; 138. [cited 2019 May 8]. Available from: http://www.==213 == .
53. Chuchu N, Takwoingi Y, Dinnes J, et al. Smartphone applications for triaging adults with skin lesions that are suspicious for melanoma. Cochrane Database Syst Rev 2018; 12:CD013192.
54. Elsner P, Bauer A, Diepgen TL, et al. Position paper: telemedicine in occupational dermatology - current status and perspectives. J der Dtsch Dermatologischen Gesellschaft (Journal Ger Soc Dermatology) 2018; 16 (8):969–974.
55. Rotemberg V, Halpern A, Dusza S, et al. The role of public challenges and data sets towards algorithm development, trust, and use in clinical practice. Semin Cutan Med Surg 2019; 38(1):E38–E42.
56. Zhang G, Yin J, Su X, et al. Augmenting multi-instance multilabel learning with sparse bayesian models for skin biopsy image analysis. Biomed Res Int 2014; 2014:305629.
57. Lim BCW, Flaherty G. Artificial intelligence in dermatology: are we there yet? Br J Dermatol. 2019; 181(1):190–191.
58. (ESR) ES of R. What the radiologist should know about artificial intelligence – an ESR white paper. Insights Imaging 2019; 10(1):44. [cited 2019 Sep 23]. Available from: https://insightsimaging.springeropen.com/articles/10.1186/s13244-019-0738-2.
59. Kohli A, Mahajan V, Seals K, et al. Concepts in U.S. Food and Drug Administration regulation of artificial intelligence for medical imaging. Am J Roentgenol 2019; 213(4):886–888. Available from: http://link.springer.com/10.1007/978-3-319-14346-0_186.