Review Article

Taiwan J Ophthalmol 2022;12: 123-129

Access this article online

Quick Response Code:

Website:
www.e-tjo.org

DOI:
10.4103/tjo.tjo_8_22

Understanding required to consider AI applications to the field of ophthalmology

Hitoshi Tabuchi[1,2*]

[1]*Department of Technology and Design Thinking for Medicine, Hiroshima University, Hiroshima, Japan,* [2]*Department of Ophthalmology, Saneikai Tsukazaki Hospital, Himeji City, Hyogo Prefecture, Japan*

***Address for correspondence:**
Dr. Hitoshi Tabuchi,
Department of Technology and Design Thinking for Medicine, Hiroshima University 1-2-3 Kasumi,
Minami-ku, Hiroshima,
Hiroshima 734-8553, Japan.
E-mail: htabuchi@hiroshima-u.ac.jp

Submission: 16-11-2021
Accepted: 28-01-2022
Published: 13-04-2022

Abstract:

Applications of artificial intelligence technology, especially deep learning, in ophthalmology research have started with the diagnosis of diabetic retinopathy and have now expanded to all areas of ophthalmology, mainly in the identification of fundus diseases such as glaucoma and age-related macular degeneration. In addition to fundus photography, optical coherence tomography is often used as an imaging device. In addition to simple binary classification, region identification (segmentation model) is used as an identification method for interpretability. Furthermore, there have been AI applications in the area of regression estimation, which is different from diagnostic identification. While expectations for deep learning AI are rising, regulatory agencies have begun issuing guidance on the medical applications of AI. The reason behind this trend is that there are a number of existing issues regarding the application of AI that need to be considered, including, but not limited to, the handling of personal information by large technology companies, the black-box issue, the flaming issue, the theory of responsibility, and issues related to improving the performance of commercially available AI. Furthermore, researchers have reported that there are a plethora of issues that simply cannot be solved by the high performance of artificial intelligence models, such as educating users and securing the communication environment, which are just a few of the necessary steps toward the actual implementation process of an AI society. Multifaceted perspectives and efforts are needed to create better ophthalmology care through AI.

Keywords:

Artificial intelligence, machine learning, medical application, ophthalmology

Introduction

The application of deep learning to ophthalmology research started with AI diagnoses of diabetic retinopathy by a group at Google, published in JAMA.[1] Our group also has reported a wide range of AI applications in ophthalmology studies from various fundus diseases such as retinal detachment,[2] glaucoma,[3] and myopic fundus disease,[4] to meibomian gland infarction,[5] lacrimal duct disease,[6] cataract surgery,[7] corneal surgery,[8] and instillation adherence.[9] Although the vast possibilities of AI deep learning applications in ophthalmology have been recognized,[10] many guidelines have been proposed by policymakers.[11] The clinical application of deep learning to ophthalmology is still in its infancy. The purpose of this review is to promote a greater understanding of the current state of deep learning as applied to research and development.

Deep Learning Led by Large IT Companies

The development boom of deep learning began when a University of Toronto team led by G. Hinton won the 2012 ImageNet Large Scale Visual Recognition Challenge (ILSVRC) by a substantial margin, leaving all other teams at 10%.[12] Currently, the world's largest AI system is YouTube's inappropriate video removal

How to cite this article: Tabuchi H. Understanding required to consider AI applications to the field of ophthalmology. Taiwan J Ophthalmol 2022;12:123-9.

system, part of Google. As of 2015, 500 h of video had been uploaded to the platform every second from all over the world; it is impossible for humans to manually review the content.[13] Improving operational efficiency through a data algorithm platform is an essential business requirement for tech companies with such a huge number of users. The social implementation of deep learning is realized in areas where a large amount of data must be processed. This contradicts the general policy of physicians, tasked with taking proper care of each and every patient. Nonprofit organizations have pointed out that it is highly likely that Google is making decisions, prioritizing its company logic without the consent of patients who have provided the images.[14] We should always keep in mind that corporations have different values from health-care professionals which they need to prioritize, at least in the short term.

Performance Limitations of Diagnostic Imaging AI

Most medical diagnostic imaging AI systems use deep learning techniques. As mentioned above, the change of the winning score of the ILSVRC, an international competition for image classification, gives a very important perspective concerning the performance limitations of medical diagnostic imaging AI. The changes in the error rate of the winning team (the lower the rate, the higher the performance) are as follows: 26% in 2011, 16.4% in 2012, 11.7% in 2013, 6.7% in 2014, 3.6% in 2015, and 3.1% in 2016.[15] Since the human classification error rate was reported to be 5%, the winning team's error rate has surpassed the humans since 2015. It was exactly in 2016 that Hinton said at the seminar that we would no longer need to train radiologists.[16] However, it has recently been noted that the performance in image classification competitions after 2015 is due to overfitting, which is one of the biggest problems with deep learning technologies. An analysis of a huge dataset to be used for classification, called ImageNet, containing 14 million images in 22,000 categories, shows that as many as 10% incorrect labels were included in the dataset, which caused overfitting through the remembering of mistakes.[17] Human classification error rates of about 5% are a performance limitation of current AI diagnostic applications,[18] which suppress unreasonable expectations for AI diagnostics. The aforementioned diabetic retinopathy diagnostic model created by training a total of 128,175 digital fundus photographs using the database already accumulated by the Google team showed 90.3% sensitivity and 98.1% specificity (AUC of 0.99) for the EyePACS-1 dataset. The performances of each AI diagnostic model created using a total of 494,661 fundus photographs by the Singapore Eye Center team were as follows: AUC of 0.958 (100% sensitivity and 91.1% specificity) for diabetic retinopathy with risk of blindness, AUC of 0.942 (96.4% sensitivity and 87.2% specificity) for glaucoma, and AUC of 0.931 (93.2% sensitivity and 88.7% specificity) for age-related macular degeneration.[19] These representative reports using large datasets and the best available AI technology provide an indication of application performance when considering the social implementation of a fundus photography AI diagnostic model.

Issues Regarding Borderline Cases

When constructing a diagnostic AI model for medical images, one encounters the problem that it is nearly impossible to perform training 100% correctly due to the many borderline cases that are difficult to classify.[20] However, a report claimed that researchers achieved a performance of 100% in optical coherence tomography (OCT) image classification applied to ophthalmology. According to a paper published in Cell by a Chinese team, researchers created a four-class classification model by training the model with 108,312°CT images obtained using the same device (SPECTRALIS OCT, HEIDELBERG ENGINEERING GmbH, Heidelberg, Germany) from multiple facilities. The four-class classification model identified choroidal neovascularization (CNV), drusen, diabetic macular edema (DME), and normal eyes (normal). The model achieved a correct response rate of 96.1%, and the highest performance was observed with the binary classification of CNV and normal, achieving AUC of 1.0 (100% sensitivity and 100% specificity).[21] The results of four-class classification by six expert doctors showed that only one doctor mistakenly identified 3 of 250 CNV images as normal, and the remaining five doctors did not misidentify any CNV images as normal. Furthermore, all six doctors did not misidentify any normal images as CNV. This means that the binary classification of normal and CNV OCT images is a task that has almost no borderline cases, even in the human eye. On the other hand, in the abovementioned four-class model AI, which test dataset contained a total of 1000 images with 250 images of each class, none of the 250 images of CNV were identified as normal. However, five CNV images were identified as DME, and three CNV images were identified as drusen. For misidentification of other diseases as CNV, seven drusen images were misidentified as CNV, and nine DME images were misidentified as CNV. This means that if CNV was defined as "positive," the sensitivity was 242/250 = 96.8%, and the specificity was 734/750 = 97.9%. Drusen and DME have sporadic cases that expert doctors would diagnose as CNV; therefore, the performance of the model decreases when the task includes such borderline cases when compared to the classifying task between CNV and normal eyes. In general, CNV and

DME, CNV and drusen in OCT are different findings for ophthalmologists. Nevertheless, in this article's diagnostic study of 250 cases per one finding, all six expert ophthalmologists confused drusen and CNV in at least two cases. Four of the six also confused DME with CNV. The discrepancy between expert ophthalmologists' diagnoses is more remarkable than between normal and CNV. In other words, it suggests that the fluctuating learning content itself is one of the factors that discourage the AI from achieving 100% performance.

Framing in AI

The frame problem[22] is a fundamental problem consisting of the impossibility for AI to learn all of the infinite conceivable possibilities. When considering a diagnostic device incorporating AI, the first thing one must do is frame "to which area it is to be applied." IDx's diabetic retinopathy diagnostic system, which became famous as the first AI diagnostic device approved by the Food and Drug Administration (FDA), has set a framing of patients who have visited a diabetes clinic to be covered by insurance.[23] AI performance fluctuates depending on which dataset is applied. For example, it was reported that an X-ray diagnostic model with a diagnostic capability of AUC 0.931 on average at two facilities showed a 10% or more reduced performance with AUC 0.815 when using X-ray images taken at other facilities.[24] Caution needs to be exercised when an AI diagnostic application is applied to an area outside the prescribed frame.

Trade-off Relationship between Interpretability and Performance

Deep learning is a machine learning (ML) model consisting of a comprehensive search by scanning each pixel of an image, and its interpretability is very low, meaning that it is hard to know what AI based its recognition of the image on. This is known as the black-box problem.[25] In spite of such problems, deep learning methods are sought due to their high performance. According to Arrieta *et al*., deep learning provides the highest accuracy and lowest interpretability among various ML models, followed by support vector machines in which humans determine image features, Bayesian models, decision trees, and linear/logistic regression; in this order, as performance decreases, interpretability increases [Figure 1]. Finally, the model that provides the lowest performance and the highest interpretability is the rule-based learning model.[26] A well-known rule-based learning model application in ophthalmology is a model for diagnosing glaucoma using the cup-to-disc ratio (C/D ratio), which represents the diameter of optic disc depression to that of the optic disc. The study was published in ophthalmology in early 2000. A glaucoma diagnosing model incorporating C/D ratio using a nerve fiber layer analyzer GDx (Carl Zeiss AG, Oberkochen, Germany) with laser light achieved a correct answer rate of 74%.[27] On the other hand, the performance of glaucoma detection using deep learning was reported in ophthalmology in 2018, achieving AUC, 0.986; sensitivity, 95.6%; and specificity, 92.0%.[28] The performance difference between the rule-based model, which provides high explainability, and the deep learning model, which has the lowest interpretability, is as much as 20%.

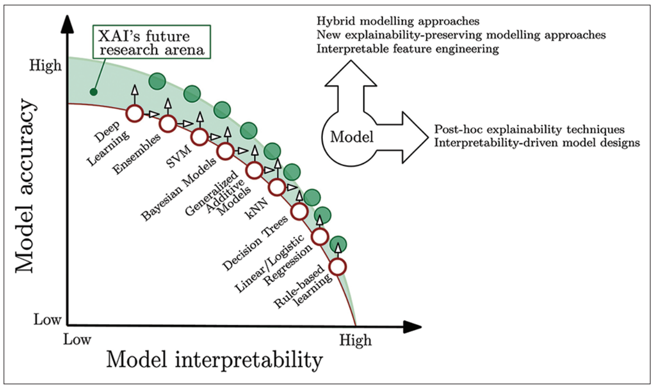

Figure 1: The trade-off between the model accuracy and the model interpretability from reference[26] (Arrieta *et al*., Information Fusion 2020[26])

Binary Classification and Segmentation Model

In binary classification with the deep learning model, it is possible to indicate the gaze point in the calculation process called the heat map by the Grad-CAM method; however, there are always multiple gaze points that are not absolute. Deep learning is the so-called black-box AI.[29] Therefore, it has become common to use the segmentation model as a simple, explainable AI model to identify the area of the lesion itself. However, the segmentation model also uses deep learning algorithms and belongs to the black-box AI category.[30] There is an important segmentation study on OCT images conducted by DeepMind, a Google company.[31] An AI model was constructed to perform fundus diagnosis by the segmentation model using a total of 15,761°CT images (3D OCT, TOPCON Corporation, Tokyo, Japan) acquired at multiple facilities in Moorfield, England. Their results showed that the error rate of 5.5% for doctors to refer their patients to the central facility was exactly the same for AI. The model provided high explainability by identifying the area with characteristic findings such as subretinal fluid and retinal pigment epithelial detachment (PED), which showed similar judgment to the diagnostic algorithm of the doctors. Compared to binary classification, age-related macular degeneration or not, for the given image, this segmentation AI is easier to understand for doctors as well as patients as it at least identifies its findings, such as PED.

Cost and Variation Problems of the Segmentation Model

The biggest problem with the segmentation model is the cost of creating annotation images (colored images) to train the AI model. In the aforementioned study by DeepMind, they had to color the areas of 15 feature findings in different colors for 877 images of the total of 15,761°CT images in the beginning stage. These colored images were used to train and construct an AI model, and the remaining 14,884°CT images were then automatically colored by the model. The paper shows one possible methodology for streamlining the labor of coloring work. It also focuses on the issue regarding model transplant to other manufacturers' OCT systems. When 527 additional segmentations of OCT images taken by another system (SPECTRALIS OCT, HEIDELBERG ENGINEERING GmbH, Heidelberg, Germany) were used to train and optimize the model, its error rate for the referral judgment was 3.7%. In an effort to realize the social implementation of AI, such methodology for improving work efficiency in the process of building an AI model is also an issue to consider. As of 2018, while the required computing power had increased more than 1000 times, the performance had not improved so much. This fact has raised a question.[32]

Compared to the classification of general objects, such as dogs, cats, and cars, the targets to be classified in medical applications include bleeding or vitiligo, which requires specialized knowledge. The differences in ability between image processing personnel known as annotators can be problematic. There is a study that investigated the inter-rater agreement for the identification range among radiological assessors evaluating honeycomb lungs.[33] Cohen's weighted κ of 43 assessors was 0.40–0.58, showing a moderate degree of agreement. Moreover, disagreement in diagnosis was observed in 29% of the actual findings. It makes sense to think that similar problems would be observed in the field of ophthalmology.

Prevalence and AI Performance

Regardless of AI diagnosis, the performance (sensitivity and specificity) of any diagnostic method and the prevalence of the diagnosis target allows prior estimation of its effectiveness under implementation. We should pay attention when targeting diseases with low prevalence particularly.

If AI diagnosis is performed for diseases with low prevalence, the false-positive rate will be extremely high no matter how high the performance of the model is, which can cause confusion in the clinical setting. For example, the prevalence of diabetic retinopathy in Japan is 1.1%,[34] and even if the highest theoretical value of AI performance with a sensitivity of 95% and specificity of 95% can be achieved in a real environment, the positive predictive value would be <10%. A common challenge in developed countries is increasing health-care costs; the economic loss due to performing unnecessary tests on too many healthy individuals cannot be ignored. I would like to introduce a paper on optic nerve papilledema published in NEJM for considering the positive predictive value.[35] This study attempted to use AI to assist nonophthalmologists with determining whether the abnormal findings of the optic disc on the fundus images are due to an ophthalmic disease or to a visual pathway lesion. The model is intended to be used by nonophthalmologists. The reported AI performance was AUC of 0.96 (95% confidence interval [CI], 0.95–0.97) with a sensitivity of 96.4% (95% CI, 93.9–98.3) and specificity of 84.7% (95% CI, 82.3–87.1). The prevalence is reported to be 9.8%, and the positive predictive value, as a result, is 39.8%.

This means that the test will indicate a positive result for papilledema in one in three people. Prevalence varies greatly depending on the clinical phase in which the AI diagnosis is performed. For example, the prevalence of brain tumors, the main cause of papilledema, has been reported to be 0.024% of emergency outpatients.[36] Not all patients with brain tumors present with papilledema.[37] Assuming that the incidence rate of the papilledema finding is roughly 50%, the prevalence of papilledema is 0.012% among 10,000 subjects. Taking into account the possibility of other rare causes and using 0.02% as the prevalence, the positive predictive value drops to merely 0.12%. In order to find one true case of papilledema, 100 healthy cases are required to be judged abnormal. In other words, if we perform fundus imaging and AI diagnosis on all patients who visited the emergency outpatient unit, it would cause confusion for the site. This AI model cannot be used in reality unless the prior probability is raised by a doctor in some way. As mentioned above, IDx's diabetes diagnosis AI can improve the positive predictive value and improve the efficiency in actual operations by limiting the target population to patients visiting the internal medicine department, or further limiting it to patients with diabetes. As mentioned earlier, the prevalence of diabetic retinopathy in Japan is 1.1%; however, the prevalence of retinopathy among diabetic patients is 15%.[38] Inevitably, the diabetes diagnosis AI can significantly increase the positive predictive value by targeting people with diabetes.

Challenges in the Social Implementation of AI

Google's team has developed a tool for diagnosing diabetic retinopathy with fundus photographs and reported on its efforts to socially implement it in multiple clinics in Thailand.[39] The fundus diagnosis AI system, which boasts an accuracy of 90% or more in an experimental environment, succeeded in interpreting more than 1000 fundus photographs a day in facilities where the implementation was successful; however, overall, the system could not diagnose more than 20% of the target patients. The researchers stated that the results were due to the network environment in Thailand, traffic access, and variations in the understanding and cooperation of inspectors at each facility. Although this system was intended to perform a diagnosis of diabetic retinopathy using fundus images in just 10 min, which could take 10 weeks in some cases, there are a huge amount of issues to be addressed before it can realize its technical assumptions. The cost and resources to reach such a point are estimated to be enormous. Any system, not only AI diagnostics, must consider that the efficiency improvement obtained under ideal conditions cannot be demonstrated when implemented in society.

Fixed Issues of AI Responsibility and AI Performance

The research paper on social implementation in Thailand by Google's team contains an important remark, stating that patients did not care whether the diagnosis was made by a doctor or by AI. This is a problem that leads to issues regarding responsibility in AI diagnosis. Recently, a view has been shared that the engineering team that built the AI model should bear responsibility.[40] In addition, the FDA requires that the performance of AI technology be fixed at the time of clinical trials.[41] That is, modifying the performance by AI feedback after marketing is not allowed. There are many disagreeing opinions on this, suggesting that only trusted architects should be allowed to make improvements to AI models.[42] This means that regulatory agencies are calling on AI architects to create an organizational culture that considers constant checking of ML processes essential.

Risk Assessment using AI

In regression estimation, unlike automatic diagnosis (identification), there is no correct answer in the first place. Therefore, the accepted range for application performance on the user side is large. It has been reported that an ML model that combines a visual field test and other clinical data showed a performance of AUC, 0.89–0.93 in a binary estimation between progressed and remained unchanged of glaucomatous visual field changes.[43] Medeiros *et al*. estimated, using AI, the average thickness of the retinal nerve fiber layer (RNFL) around the optic disc from OCT images and fundus photographs of the optic nerve head taken with a nonmydriatic stereo fundus camera.[44] The results showed that the actual mean RNFL thickness by OCT was 82.5 ± 16.8 µm, the estimated mean RNFL thickness by AI was 83.3 ± 14.5 µm, and the correlation coefficient of Pearson was r = 0.832. These results indicated a very accurate AI estimation model. There has been another study that applied this method to predict the risk of glaucoma progression using fundus photographs only.[45] According to the study, the mean RNFL of the papilla estimated by AI was 88.7 ± 9.4 mm (mean observation period: 4.4 ± 3.8 years) in the group that progressed to glaucoma while it was 92.1 ± 7.2 mm (mean observation period: 6.3 ± 3.7 years) in the group that did not progress to glaucoma, showing a significant difference ($P < 0.001$). Every time AI estimated the mean RNFL to be 10 µm thinner than the baseline, the risk of developing glaucoma increased by 56%, and when the annual decrease in the mean RNFL estimated by AI was >1 µm, the glaucoma risk also increased by 99%. The level of increase when determining the risk of visual field progression used in actual practice is 108%, which is comparable to the AI estimation performance.

Conclusion

The evaluation of medical applications of AI starts after it is implemented in society. Medical artificial intelligence has the characteristics of software services that are thought to be best for lean startup management,[46] which run a cycle of quick feedback and improvement after release. Feedback from actual clinical practice should be utilized to improve the performance of AI. The FDA, however, does not allow such methodology at this time, and it is not absolutely clear if feedback-based performance improvements will be the best solution.

In an effort to determine which area to apply AI, what kind of legal framework should be established, and the best solutions for intricately complex problems, the wisdom of all ophthalmologists around the world must be brought together to provide better ophthalmological care in future.

Financial support and sponsorship

Nil.

Conflicts of interest

The authors declare that there are no conflicts of interests of this paper.

References

1. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, *et al*. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA 2016;316:2402-10.
2. Ohsugi H, Tabuchi H, Enno H, Ishitobi N. Accuracy of deep learning, a machine-learning technology, using ultra-wide-field fundus ophthalmoscopy for detecting rhegmatogenous retinal detachment. Sci Rep 2017;7:9425.
3. Masumoto H, Tabuchi H, Nakakura S, Ishitobi N, Miki M, Enno H. Deep-learning classifier with an ultrawide-field scanning laser ophthalmoscope detects glaucoma visual field severity. J Glaucoma 2018;27:647-52.
4. Sogawa T, Tabuchi H, Nagasato D, Masumoto H, Ikuno Y, Ohsugi H, *et al.* Accuracy of a deep convolutional neural network in the detection of myopic macular diseases using swept-source optical coherence tomography. PLoS One 2020;15:e0227240.
5. Maruoka S, Tabuchi H, Nagasato D, Masumoto H, Chikama T, Kawai A, *et al.* Deep Neural network-based method for detecting obstructive meibomian gland dysfunction with *in vivo* laser confocal microscopy. Cornea 2020;39:720-5.
6. Imamura H, Tabuchi H, Nagasato D, Masumoto H, Baba H, Furukawa H, *et al*. Automatic screening of tear meniscus from lacrimal duct obstructions using anterior segment optical coherence tomography images by deep learning. Graefes Arch Clin Exp Ophthalmol 2021;259:1569-77.
7. Morita S, Tabuchi H, Masumoto H, Yamauchi T, Kamiura N. Real-time extraction of important surgical phases in cataract surgery videos. Sci Rep 2019;9:16590.
8. Hayashi T, Masumoto H, Tabuchi H, Ishitobi N, Tanabe M, Grün M, *et al.* A deep learning approach for successful big-bubble formation prediction in deep anterior lamellar keratoplasty. Sci Rep 2021;11:18559.
9. Nishimura K, Tabuchi H, Nakakura S, Nakatani Y, Yorihiro A, Hasegawa S, *et al.* Evaluation of automatic monitoring of instillation adherence using eye dropper bottle sensor and deep learning in patients with glaucoma. Transl Vis Sci Technol 2019;8:55.5.
10. Balyen L, Peto T. Promising artificial intelligence-machine learning-deep learning algorithms in ophthalmology. Asia Pac J Ophthalmol (Phila) 2019;8:264-72.
11. Campbell JP, Lee AY, Abràmoff M, Keane PA, Ting DS, Lum F, *et al*. Reporting guidelines for artificial intelligence in medical research. Ophthalmology 2020;127:1596-9.
12. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521:436-44.
13. Iqbal M. YouTube Revenue and Usage Statistics; 2021. Available from: https://www.businessofapps.com/data/YouTube-statistics/. [Last accessed on 2021 May 16].
14. Blumenthal D. Why Google's Move into Patient Information Is a Big Deal: Technology and Analytics. Harvard Business Review; 2019. Available from: https://hbr.org/2019/11/why-googles-move-into-patient-information-is-a-big-deal. [Last accessed on 2021 Oct 15].
15. ImageNet Large Scale Visual Recognition Challenge (ILSVRC): Competition. Available from: https://www.image-net.org/challenges/LSVRC/. [Last accessed on 2021 Oct 15].
16. Hinton G. On Radiology; 2016. Available from: https://www.youtube.com/watch?v=2HMPRXstSvQ. [Last accessed on 2021 Oct 15].
17. Beyer L, Hénaff OJ, Kolesnikov A, Zhai X, van den Oord A. Are We Done with ImageNet?; 2020. Available from: https://arxiv.org/abs/2006.07159v1. [Last accessed on 2021 Oct 15].
18. Rayner LO. AI Competitions Don't Produce Useful Models; 2019. Available from: https://lukeoakdenrayner.wordpress.com/2019/09/19/ai-competitions-dont-produce-useful-models/. [Last accessed on 2021 Oct 17].
19. Ting DS, Cheung CY, Lim G, Tan GS, Quang ND, Gan A, *et al.* Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. JAMA 2017;318:2211-23.
20. Hilbig N. How Far can artificial Intelligence Go? The 8 Limits of Machine Learning. Available from: https://medium.com/codex/how-far-can-artificial-intelligence-go-the-8-limits-of-machine-learning-383dd9b2f7bd. [Last accessed on 2021 Oct 16].
21. Kermany DS, Goldbaum M, Cai W, Valentim CC, Liang H, Baxter SL, *et al.* Identifying medical diagnoses and treatable diseases by image-based deep learning. Cell 2018;172:1122-31.e9.
22. McCarthy J, Hayes PJ. Some philosophical problems from the standpoint of artificial intelligence. In: Meltzer B, Michie D, editors. Machine Intelligence. Edinburgh, UK: Edinburgh University Press; 1969;4:463-502.
23. van der Heijden AA, Abramoff MD, Verbraak F, van Hecke MV, Liem A, Nijpels G. Validation of automated screening for referable diabetic retinopathy with the IDx-DR device in the Hoorn Diabetes Care System. Acta Ophthalmol 2018;96:63-8.
24. Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. PLoS Med 2018;15:e1002683.
25. Dickson B, The Dangers of Trusting Black-Box Machine Learning; 2020. Available from: https://bdtechtalks.com/2020/07/27/black-box-ai-models/. [Last accessed on 2021 Oct 17].
26. Arrieta AB, Díaz-Rodríguez N, del Sera J, Bennetot A, Tabik S, Barbado A, *et al*. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Inf Fus 2020;58:82-115.
27. Choplin NT, Lundy DC. The sensitivity and specificity of scanning laser polarimetry in the detection of glaucoma in a clinical setting. Ophthalmology 2001;108:899-904.
28. Li Z, He Y, Keel S, Meng W, Chang RT, He M. Efficacy of a deep learning system for detecting glaucomatous optic neuropathy based on color fundus photographs. Ophthalmology 2018;125:1199-206.
29. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual explanations from deep networks via gradient-based localization. Int J Comput Vis 2020;128:336-59.
30. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell 2019;1:206-15.
31. De Fauw J, Ledsam JR, Romera-Paredes B, Nikolov S, Tomasev N, Blackwell S, *et al.* Clinically applicable deep learning for diagnosis and referral in retinal disease. Nat Med 2018;24:1342-50.
32. Piekniewski's Blog. "AI Winter Is Well On Its Way"; 2018. Available from: https://blog.piekniewski.info/2018/05/28/ai-winter-is-well-on-its-way/. [Last accessed on 2021 Oct 17].
33. Watadani T, Sakai F, Johkoh T, Noma S, Akira M, Fujimoto K, *et al.* Interobserver variability in the CT assessment of honeycombing in the lungs. Radiology 2013;266:936-44.
34. Yasuda M. Epidemiology of diabetic retinopathy. OCULISTA (Japanese) 2013;8:1-5.
35. Milea D, Najjar RP, Zhubo J, Ting D, Vasseneix C, Xu X, *et al.* Artificial intelligence to detect papilledema from ocular fundus photographs. N Engl J Med 2020;382:1687-95.
36. Comelli I, Lippi G, Campana V, Servadei F, Cervellin G. Clinical presentation and epidemiology of brain tumors firstly diagnosed in adults in the Emergency Department: A 10-year, single center retrospective study. Ann Transl Med 2017;5:269.
37. Friedman DI. In: Miller NR, Walsh FB, Hoyt WF, editors. Papilledema, Walsh and Hoyt's Clinical Neuro-ophthalmology. 6th ed., Vol. 1. Baltimore, MD: Lippincott Williams & Wilkins; 2005. p. 237-92.
38. Yasuda M, Kiyohara Y, Wang JJ, Arakawa S, Yonemoto K, Doi Y,

et al. High serum bilirubin levels and diabetic retinopathy: the Hisayama Study. Ophthalmology 2011;118:1423-8.

39. Beede E, Baylor E, Hersch F, Iurchenko A, Wilcox L, Ruamviboonsuk P, *et al.* A Human- Centered Evaluation of a Deep Learning System Deployed in Clinics for the Detection of Diabetic Retinopathy. Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. Available from: https://dl.acm.org/doi/abs/10.1145/3313831.3376718. [Last accessed on 2021 Oct 17].
40. Babic B, Cohen IG, Evgeniou T, Gerke S. When Machine Learning Goes Off the Rails. Guide to Managing the Risks. Harvard Business Review; 2021. Available from: https://hbr.org/2021/01/when-machine-learning-goes-off-the-rails. [Last acessed on 2021 Oct 17].
41. ICMR. Informal Innovation Network. Horizon Scanning Assessment Report – Artificial Intelligence; 2021. Available from: http://www.icmra.info/drupal/sites/default/files/2021-08/horizon_scanning_report_artificial_intelligence.pdf. [Last accessed on 2021 Sep 01].
42. FDA. Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) Action Plan; 2021. Available from: https://www.fda.gov/media/145022/download?utm_medium=email&utm_source=govdelivery. [Last accessed on 2021 Sep 01].
43. Dixit A, Yohannan J, Boland MV. Assessing glaucoma progression using machine learning trained on longitudinal visual field and clinical data. Ophthalmology 2021;128:1016-26.
44. Medeiros FA, Jammal AA, Thompson AC. From machine to machine: An OCT-trained deep learning algorithm for objective quantification of glaucomatous damage in fundus photographs. Ophthalmology 2019;126:513-21.
45. Lee T, Jammal AA, Mariottoni EB, Medeiros FA. Predicting glaucoma development with longitudinal deep learning predictions from fundus photographs. Am J Ophthalmol 2021;225:86-94.
46. Ries E. The Lean Startup. New York, NY: Crown Publishing; 2011.