Editorial

Deep Learning in Medicine. Are We Ready?

Daniel SW Ting, [1,2]*MD, PhD*, Tyler H Rim,[3]*MD, MBA*, Yoon Seong Choi, [4]*MD, PhD*, Joseph R Ledsam, [5]*MBChB*

The real-world application of artificial intelligence (AI), machine learning (ML) and deep learning (DL), have generated significant interest throughout the computer science and medical communities in recent years. This interest has been accompanied by no small amount of hype. Though the term 'ML' was coined 50 years ago by Arthur Samuel, who stated that machines should have the ability to learn without being programmed,[1] the advent of the graphics processing unit (GPU) has enabled much improved processing power and enabled new possibilities with AI. DL—an approach that utilises multiple neural networks to learn representation of data using multiple levels of abstraction[2]—has revolutionised the computer vision field, and achieved substantial jumps in diagnostic performance for image recognition, speech recognition, and natural language processing.[2] In the technical world, DL has been heavily used in autonomous vehicles,[3] gaming[4,5] and numerous smart phone applications. The availability of different software (e.g. Caffe, Tensorflow), and the off-the-shelf convolutional neural networks (e.g. AlexNet, VGGNet, ResNet and GoogleNet) have removed barriers to entry for many academics and clinicians, resulting in the recent surge of interest within the medical settings. To date, this technique has shown promising diagnostic performance, across specialties including ophthalmology (e.g. detection of diabetic retinopathy [DR], glaucoma and age-related macular degeneration from fundus photographs and optical coherence tomographs),[6-11] radiology (e.g. detection of tuberculosis from chest X-rays [CXRs], intracranial haemorrhage from computed tomography of the brain),[12-15] and dermatology (e.g. detection of malignant melanoma from skin photographs).[16]

DL is a tool that can, when applied effectively, serve multiple roles in different medical settings. Examples of this include screening, triaging referral urgency, prognosticating and monitoring diseases progression. In order to increase the explainability of the outcome, many of the more recent DL systems experiment with attempts to visualise the decision process. Examples include demonstrating disease activities via heat maps,[10] and displaying pathological features with image segmentation. With such abilities, this may help to increase the DL systems' adoption rate by physicians and their acceptability by patients.

In ophthalmology, one of the most promising areas is DR screening. Globally, 600 million people will have diabetes by 2040; a third will develop DR.[17] Given the increasing prevalence of diabetes and ageing population, DR screening programmes are constantly challenged by issues related to implementation, availability of human assessors and long-term financial sustainability.[18] In order to rectify the manpower shortage, DL systems can be an alternative DR screening tool. In 2016, both Gulshan et al and Abramoff et al reported excellent diagnostic performances of the DL systems in detecting referable DR using publicly available datasets, with area under the receivers' operating curves (ROCs) (AUC) of >0.95 in both studies.[8,19] Ting and co-workers have also developed and tested a DL system for identifying DR, and related eye diseases using nearly half a million images from multiethnic community, population-based and clinical datasets.[7] Consistent with the minimum screening performance (sensitivity of at least 80%) set by the Diabetes United Kingdom,[20] the diagnostic performance of this DL system was clinically acceptable with AUC of >90%, sensitivity of >90% and specificities >85% for referable DR, vision-threatening DR, glaucoma suspect and age-related macular degeneration. More importantly, this DL system was also tested on 10 external datasets, consisting of multiple ethnicities and settings (by patients' demographics and glycaemic control, status of pupil dilation, retinal cameras and width of field for retinal images), using diverse reference standards in DR assessment by professional graders, optometrists or retinal specialists. In order to ensure generalisability, it is always important to test a DL system on previously unseen datasets. A similar

[1]Singapore Eye Research Institute, Singapore National Eye Centre, Singapore
[2]Duke-NUS Medical School, National University of Singapore, Singapore
[3]Department of Ophthalmology, Institute of Vision Research, Yonsei University College of Medicine, Seoul, Korea
[4]Department of Radiology, Research Institute of Radiological Science, Severance Hospital, Yonsei University College of Medicine, Seoul, Korea
[5]DeepMind, London, United Kingdom
Address for Correspondence: Asst/Prof Daniel Ting Shu Wei, Singapore Eye Research Institute, Singapore National Eye Centre, 11 Third Hospital Avenue, Singapore 168751.
Email: daniel.ting.s.w@singhealth.com.sg

example was demonstrated by Abramoff et al in a recent United States Food and Drug Administration (US FDA)-approved autonomous DR detection software – IDx, which was tested in prospective clinical trials in the US.[21] Thus, DL systems for DR can potentially be deployed in the countries with and without existing DR screening programmes, as semi-automated or fully-automated models, with the aim to prevent DR-related visual impairment for the global population with diabetes worldwide.

Skin cancer is another major public health concern.[22] In the US, it is estimated that approximately 9000 people are dying from melanoma each year, with $3.3 billion of skin cancer treatment costs attributable to skin melanoma.[23] Given the shortage of dermatologists, a DL system may be an alternative solution for this. Esteva et al reported a robust, dermatologist-level comparable DL system for detection of skin cancer.[16] Using a dataset of 129,450 clinical images with 2032 different diseases, this DL system was tested against 21 board-certified dermatologists on biopsy-proven clinical images (photographic and dermoscopic images) for 2 groups – keratinocyte carcinomas (the most common cancer) versus benign seborrhoeic keratosis; and malignant melanomas (the deadliest skin cancer) versus benign nevi. This DL system showed on par diagnostic performance with all tested dermatologists, with AUC of >0.90 for keratinocyte carcinoma (skin photographs) and melanoma (skin photographs and dermoscopic images). Future research is beneficial to assess the cost-effectiveness of this DL system for patients with skin lesions.

Pulmonary tuberculosis (TB) is an infectious disease that poses a significant public health problem, causing 1.5 million deaths worldwide in 2013.[24] CXRs play an important role in screening and diagnosis of pulmonary TB, but their interpretation requires radiological expertise and is resource-intensive, particularly in developing countries. As such, there has been interest in the development of effective automated DL methods for detection and diagnosis of pulmonary TB from CXRs. Both Lakhani et al and Hwang et al have reported good diagnostic performance in using DL systems for detection of TB.[12,13] Using AlexNet and GoogLeNet, Lakhani and co-worker reported an AUC of 0.99 in detection of TB in a dataset consisting of approximately 1000 CXRs.[12] The testing dataset for this study, however, may be underpowered.[25] Using a much larger sample size (approximately 60,000 CXRs), Hwang et al, recently, also reported a robust DL system to detect TB (AUC = 0.988), with the ability to localise abnormal lesions (AUC = 0.977). The reference standard consists of 15 readers—5 non-radiology physicians, 5 general radiologists and 5 thoracic radiologists. This robust performance, again, showed consistency in 6 external datasets (4 Korean datasets, 1 US dataset, and 1 Chinese dataset), with AUC of >0.97. This study is a good example to emphasise the importance of having multiple reference standards, independent datasets and the ability to localise the disease activity areas. The algorithm published in this paper can be tested via the website, https://insight.lunit.io/.

Aside from screening, the DL system has been reported to be a robust tool to triage the urgency of referrals to the tertiary healthcare settings. Earlier this year, DeepMind and Moorfields Eye Hospital delineated 15 different retinal morphologic features from retinal optical coherence tomography scans, using a 2-stage convolutional neural network (CNN) architecture consisting of separate segmentation and classification networks. This DL system has excellent ability (AUC >0.90) to make a referral triage decision from 4 categories (urgent, semi-urgent, routine, observation), and classifies the presence of 10 different retinal diseases.[26] This DL system may be a useful clinical tool to be implemented in the rapid access "virtual" clinics that are now widely used for triaging of macular disease in the United Kingdom.[27]

In this issue, a review by Liew et al describes the role of AI in radiology with a focus on the Singapore setting.[28] AI expands beyond helping or substituting human work, to extracting quantitative information for clinical decision-making and prognosis prediction; the authors provide a comprehensive commentary of the willingness of local radiologists to work together and collaborate with key stakeholders within the context of our "smart nation". This timely review and its declaration of intention to embrace the science and implementation of new tools is a laudable first step, and should perhaps also be expanded to take into account other promising techniques being added to the toolkit of diagnostic imaging to contribute to precision medicine. In radiology, an emerging technique is the so-called "radiomics", where high-dimensional numeric information is extracted from the medical image and put into ML to non-invasively predict relevant clinical information. For example, glioma (the most common primary tumour) and glioblastoma (a grade IV glioma and the most malignant glioma), have poor prognosis with median survival of only 18 months, making early correct diagnosis and prognosis prediction important. It has been reported that radiomics showed excellent performance (AUC >0.90) in preoperatively differentiating confusing cases of glioblastoma and primary central nervous system lymphoma (PCNSL) which may show similar magnetic resonance imaging (MRI) findings but have different treatment strategies.[29] Bae et al reported that radiomics can improve prognosis of glioblastoma beyond the established prognostic factors including clinical and molecular subtype information. In this study, when radiomics models were trained with MRI-based radiomic features using random

survival forest on training cohort (n = 163) and integrated into clinical and molecular information, the prognostication improved (integrated area under the time-dependent ROC curve showed improved performance on the test set [n = 54], integrated area under the time-dependent ROC curve, 0.696 vs 0.782, $P = 0.04$) for overall survival prediction.[30] Radiomics is recently evolving from extracting handcrafted features based on specific equations, to automatically extract and train the algorithm by CNN. Chang et al reported that residual network can predict isocitrate dehydrogenase in grade II to IV gliomas, a major molecular subtype for treatment response and prognosis.[31] In this study, multiplanar preoperative MR images were put into the 34-layer residual network and trained, validated and tested on a total of 496 multicentre patients, yielding excellent performance (AUC = 0.94, accuracy = 95.7%).

DL methods may be employed to predict features and pathologies beyond those conventionally used in clinical practice. Poplin et al recently reported an interesting DL system used to predict cardiovascular risk factors (e.g. age, gender, blood pressure) from fundus photographs.[32,33] Along with convolutional neural network, the recurrent neural network can be applied for natural language processing, analysing the longitudinal medical record to predict in-hospital mortality, 30-day unplanned readmission, prolonged length of stay, and all of a patient's final discharge diagnoses.[34]

Although DL systems have been reported to have robust performances in different clinical settings, many limitations still exist in the literature in terms of safe integration into practice. For example, there have been many studies investigating plain film X-ray, but most are limited to a binary classification between normal and one disease or grade in a certain disease. Where studies investigate larger numbers of disease classes, reported accuracy tends to be lower.[35] The imageability (also known as gradability) also remains as a challenging aspect in any DL system. Most studies have trained and validated with good quality photographs, yielding robust diagnostic performance (AUC >0.90). One given algorithm can yield variable performance, depending on the quality of input data. For example, MRI can vary according to the scan protocol or vendor manufacturer (e.g. 1.5T vs 3T, or Philips vs Siemens), thus affecting the performance. One also needs to be mindful about the concept of "garbage in and garbage out". In other words, even in the most robust convolutional neural network, the accuracy of the teaching datasets ground truth is perhaps the most important consideration in a study. To assure reproducibility and generalisability, larger training sample size, validation on more variable study cohort, and sharing details and even codes of preprocessing and training algorithm are mandatory.[10] As such, the radiomic quality score (RQS) system has been published to measure the quality of radiomics study, allowing the description of the details of image processing pipeline, training algorithms, the characteristics, and inclusion/exclusion criteria of the study cohort.[36]

In summary, AI using DL is a promising novel state-of-art technology for the medical world. And it is crucial that we, as a community, ensure a robust training datasets with reliable ground truths and to test the implementation of these models in clinical practice. The formation of the Radiological AI, Data Science and Imaging Informatics (RADII) under the Singapore Radiological Society is a good platform to gather all stakeholders from the clinical and ML community.[28] Although there are still many challenges that need to be solved prior to the mass AI adoption in healthcare, it is important for physicians to collaborate widely,[37] aiming to improve the work efficiency and the access to tertiary health, from Singapore, and potentially to the global setting.

Acknowledgement

Financial Disclosure: Dr Daniel Ting is the co-inventor of a deep learning system for retinal diseases, the co-founder and shareholder of EyRIS Ptd Ltd, Singapore. Dr Tyler Rim is a shareholder of Mediwhale, Korea. Dr Joseph Ledsam is a clinician scientist and employee of DeepMind, UK.

REFERENCES

1. Samuel AL. Some Studies in Machine Learning using the Game of Checkers. In: Computer Games. New York: Springer;1988. p. 335-65.
2. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521:436-44.
3. Chen C, Seff A, Kornhauser A, Xiao J. DeepDriving: Learning Affordance for Direct Perception in Autonomous Driving. Proceedings of the IEEE International Conference on Computer Vision (ICCV) 2015. p. 2722-30.
4. Silver D, Huang A, Maddison CJ, Guez A, Sifre L, van den Driessche G, et al. Mastering the game of Go with deep neural networks and tree search. Nature 2016;529:484-9.
5. Silver D, Schrittwieser J, Simonyan K, Antonoglou I, Huang A, Guez A, et al. Mastering the game of Go without human knowledge. Nature 2017;550:354-9.
6. Ting DSW, Liu Y, Burlina P, Xu X, Bressler NM, Wong TY. AI for medical imaging goes deep. Nat Med 2018;24:539-40.
7. Ting DSW, Cheung CY, Lim G, Tan GSW, 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.
8. 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.
9. Ting DSW, Wu WC, Toth C. Deep learning for retinopathy of prematurity screening. Br J Ophthalmol 2018. pii: bjophthalmol-2018-313290.

10. Ting DSW, Pasquale LR, Peng L, Campbell JP, Lee AY, Raman R, et al. Artificial intelligence and deep learning in ophthalmology. Br J Ophthalmol 2019;103:167-75.
11. 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.
12. Lakhani P, Sundaram B. Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks. Radiology 2017;284:574-82.
13. Hwang EJ, Park S, Jin KN, Kim JI, Choi SY, Lee JH, et al. Development and validation of a deep learning-based automatic detection algorithm for active pulmonary tuberculosis on chest radiographs. Clin Infect Dis 2018.
14. Ting DSW, Tan TE, Lim CCT. Development and validation of a deep learning system for detection of active pulmonary tuberculosis on chest radiographs: clinical and technical considerations. Clin Infect Dis 2018.
15. Titano JJ, Badgeley M, Schefflein J, Pain M, Su A, Cai M, et al. Automated deep-neural-network surveillance of cranial images for acute neurologic events. Nat Med 2018;24:1337-41.
16. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017;542:115-8.
17. Yau JW, Rogers SL, Kawasaki R, Lamoureux EL, Kowalski JW, Bek T, et al. Global prevalence and major risk factors of diabetic retinopathy. Diabetes Care 2012;35:556-64.
18. Ting DS, Cheung GC, Wong TY. Diabetic retinopathy: global prevalence, major risk factors, screening practices and public health challenges: a review. Clin Exp Ophthalmol 2016;44:260-77.
19. Abramoff MD, Lou Y, Erginay A, Clarida W, Amelon R, Folk JC, et al. Improved automated detection of diabetic retinopathy on a publicly available dataset through integration of deep learning. Invest Ophthalmol Vis Sci 2016;57:5200-6.
20. Yang Z, Chan YS, Wong TM. Effects of kainic acid administered to the caudal ventrolateral medulla on arterial blood pressure in the spontaneously hypertensive and normotensive Wistar-Kyoto rats. Neurosci Lett 1996;202:145-8.
21. Abramoff MD, Lavin PT, Birch M, Shah N, Folk JC. Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. NPJ Digital Medicine 2018;39:1-8.
22. Apalla Z, Lallas A, Sotiriou E, Lazaridou E, Ioannides D. Epidemiological trends in skin cancer. Dermatol Pract Concept 2017;7:1-6.
23. Giuffre G, Lodato G, Dardanoni G. Prevalence and risk factors of diabetic retinopathy in adult and elderly subjects: the Casteldaccia eye study. Graefes Arch Clin Exp Ophthalmol 2004;242:535-40.
24. Dheda K. Getting bang for buck in the latent tuberculosis care cascade. Lancet Infect Dis 2016;16:1209-10.
25. Ting DS, Yi P, Hui F. Clinical applicability of deep learning system in detecting tuberculosis using chest radiography. Radiology 2018;286:729-31.
26. 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.
27. Buchan JC, Amoaku W, Barnes B, Cassels-Brown A, Chang BY, Harcourt J, et al. How to defuse a demographic time bomb: the way forward? Eye (Lond) 2017;31:1519-22.
28. Liew CJY, Krishnaswamy P, Cheng LTE, Tan CH, Poh ACC, Lim TCC. Artificial intelligence and radiology in Singapore: championing a new age of augmented imaging for unsurpassed patient care. Ann Acad Med Singapore 2019;48:15-23.
29. Suh HB, Choi YS, Bae S, Ahn SS, Chang JH, Kang SG, et al. Primary central nervous system lymphoma and atypical glioblastoma: differentiation using radiomics approach. Eur Radiol 2018;28:3832-9.
30. Bae S, Choi YS, Ahn SS, Chang JH, Kang SG, Kim EH, et al. Radiomic MRI phenotyping of glioblastoma: improving survival prediction. Radiology 2018;289:797-806.
31. Chang K, Bai HX, Zhou H, Su C, Bi WL, Agbodza E, et al. Residual convolutional neural network for the determination of IDH status in low-and high-grade gliomas from MR imaging. Clin Cancer Res 2018;24:1073-81.
32. Poplin R, Varadarajan AV, Blumer K, Liu Y, McConnell MV, Corrado GS, et al. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. Nature Biomedical Engineering 2018;2:158-64.
33. Ting DS, Wong TY. Eyeing cardiovascular risk factors. Nature Biomedical Engineering 2018;2:140-1.
34. Rajkomar A, Oren E, Chen K, Dai AM, Hajaj N, Hardt M, et al. Scalable and accurate deep learning with electronic health records. NPJ Digital Medicine 2018;1:18.
35. Choi JY, Yoo TK, Seo JG, Kwak J, Um TT, Rim TH. Multi-categorical deep learning neural network to classify retinal images: a pilot study employing small database. PLoS One 2017;12:e0187336.
36. Lambin P, Leijenaar RT, Deist TM, Peerlings J, de Jong EEC, van Timmeren J, et al. Radiomics: the bridge between medical imaging and personalized medicine. Nat Rev Clin Oncol 2017;14:749.
37. Yi PH, Hui F, Ting DS. Artificial intelligence and radiology: collaboration is key. J Am Coll Radiol 2018;S1546-1440:30001-2.