Use of artificial intelligence in skin cancer diagnosis and management

The challenge now is how to implement artificial intelligence technology safely into clinical practice

Artificial intelligence is a branch of computer science that, in broad terms, deals with either decision making or classification. The aim of artificial intelligence is to surpass human cognitive functioning such that automated decisions can be made. Machine learning — an application of artificial intelligence — is commonly used in image recognition. In general, the machine, or algorithm, learns from exposure to a large dataset. Once learning has taken place, the algorithm can be applied to unseen data. The potential advantages of this approach in health care are clear: machines can learn from very large datasets in relatively short time frames and can apply themselves to new data without fatigue or intra-observer replication error.

Machine learning has recently demonstrated remarkable performance in image-based diagnosis across various medical fields, including ophthalmology, radiology, pathology and dermatology. In dermatology, the primary focus has been on developing machine learning systems that facilitate classification and decision support for skin cancer management. Skin cancer (including melanocytic and keratinocytic malignancy) is the most common cancer in Australia and among Caucasian populations worldwide. Melanoma is responsible for the majority of skin cancer deaths in Australia and has various presentations.[1,2] While dermoscopy has improved the accuracy of melanoma diagnosis, significant variability occurs and is largely a function of clinical expertise. Recent studies show that machine learning algorithms have the potential to surpass the diagnostic performance of experts, and the challenge now is how to implement this new technology safely into clinical practice.

Although there are a number of machine learning algorithms that could be used in the dermatology setting, convolutional neural networks (CNNs) are the most promising. This is largely because they learn from data without any feature specification, and they are known to exhibit superior performance for image recognition in comparison with other machine learning algorithms.[3] The aim of the CNN is to generalise its previously learned knowledge on unseen images beyond the training dataset. There are numerous parameters within a CNN that can be tweaked to maximise algorithm performance. Most of these parameters are adjusted automatically by the algorithm, without user input. Therefore, very little can be known, in principle, about why and how the algorithm reaches any particular decision. Currently, there are efforts underway to reduce the "black box" effect of CNNs. Some commercial software programs coupled to imaging devices will provide the user with comparable lesions to justify the algorithm's output and improve transparency. However, this retrieval system may fail for rare or unseen cases and does not provide a decision-making process. While the black box phenomenon remains, there are two potentially negative implications for clinical practice: first, clinicians may have difficulty upskilling by following the algorithms' outputs; and second, there exists the potential for deskilling and underperforming due to an over-reliance on technology.[4,5] The effect of a faulty system has been explored by manipulating a previously trusted algorithm to generate incorrect classifications and found that doctors of all experience levels were susceptible to being misled by the recommendation.[5]

Algorithm performance is dependent on both the size and quality of the training image dataset and on whether the algorithm is used in situations for which it was intended. Depending on the training set, the device may be limited in its ability to diagnose specific lesions (eg, non-pigmented), or lesions in certain skin types (eg, darker skin) or sites (eg, scalp or acral). Retrospective image databases used to train algorithms may be associated with bias. In addition, artefacts (eg, hair, dermoscopic gel, air bubbles, rulers, pen markings, reflections) can distract from key features. However, if a CNN is trained on a large enough cohort, it can learn to deal with potential artefacts. Nonetheless, unbiased lesion selection and standardised image capture would invariably improve algorithm performance, and recent advances in three-dimensional (3D) imaging modalities will enable this.[6]

Several studies have now shown that CNNs trained on retrospective image data collected at a single time point are capable of classifying skin cancer with sensitivities and specificities equal or superior to that of dermatologists (Box 1),[5,7–9,11] and clinicians with less experience gain most from AI support under experimental conditions.[5] Hypomelanotic and acral melanoma can be more challenging to diagnose clinically,[1] and this could potentially present a challenge for automated classification. However, CNNs have achieved greater accuracy for hypopigmented and acral lesions in comparison with human experts, at least in silica.[9,11] In addition to clinical images, CNNs have been applied to histopathological images of melanoma and benign naevi with promising results.[10]

The ground truth for lesion diagnosis

The gold standard for melanoma diagnosis is histopathological assessment. However, there exists significant inter- and intra-observer variability in histological diagnostic labels attributed to atypical

Miki Wada[1]
ZongYuan Ge[1]
Stephen J Gilmore[2]
Victoria J Mar[1,3]

1 Monash University, Melbourne, VIC.
2 Skin Health Institute, Melbourne, VIC.
3 Victorian Melanoma Service, Alfred Hospital, Melbourne, VIC.

victoria.mar@monash.edu

doi: 10.5694/mja2.50759

1 Comparison of skin cancer classification tasks by artificial intelligence (AI) systems and dermatologists/pathologists

Study	AI architecture	Images	Classification task	Training dataset size	Test dataset size	AI			Dermatologists/pathologists		
						Sensitivity	Specificity	AUC/overall accuracy	Sensitivity	Specificity	AUC/overall accuracy
Tschandl[5]	ResNet34 CNN	Clinical (dermoscopic)	Benign *v* malignant *v* non-neoplastic skin lesions	10 015	1412	0.81 (0.79–0.83)*	0.92 (0.90–0.93)*	0.73† (0.70–0.76)*	0.80 (0.78–0.83)*	0.80 (0.77-0.82)*	0.60† (0.57–0.63)*,‡
									0.86 (0.84–0.88§)*	0.88 (0.87–0.90§)*	0.74† (0.71–0.77§)*
Esteva[7]	GoogleNet Inception v3 CNN	Clinical (macroscopic, dermoscopic)	Benign *v* malignant *v* non-neoplastic skin lesions	129 450	1942	na	na	72.1%¶ ± 0.9%	na	na	66.0%¶
Haenssle[8]	GoogleNet Inception v4 CNN	Clinical (macroscopic, dermoscopic)	Benign melanocytic naevi *v* melanoma	> 100 000	100	86.6%**	82.5%**	0.86**	86.6%**	71.3%**	0.79**
						88.9%††	82.5%††	0.86††	88.9%††	75.7%††	0.82††
Tschandl[9]	GoogleNet Inception v3 CNN	Clinical (macroscopic, dermoscopic)	Benign *v* malignant hypo-pigmented lesions	13 724	2072	81%	53.5%	0.73	78%	51.3%	0.68
Hekler[10]	ResNet50 CNN	Histopathology	Benign naevus *v* melanoma	595	100	76%	60%	na	51.8%‡‡	66.5%‡‡	na
Fujisawa[11]	GoogleLeNet DCNN	Clinical (macroscopic)	Benign *v* malignant skin lesions§§	4867	1142	96.3%	89.5%	92.4%¶ ± 2.1%	na	na	85.3%¶ ± 3.7%

AUC = area under the curve; na = not applicable. * 95% CI. † Youden statistic. ‡ Clinicians with varied experience and training. § Clinician accuracy with multiclass probabilistic AI support. ¶ Overall accuracy. ** Level I: AI and human readers provided with dermoscopic images only. †† Level II: AI provided with dermoscopic images only, human readers provided with dermoscopic images, macroscopic images and additional clinical information. ‡‡ Pathologist. §§ 52.6% of melanomas in this study were acral. ◆

melanocytic lesions.[12] The existence of such variability in diagnoses poses the dilemma of whether the CNN has learnt from the correct set of diagnoses. Consensus diagnoses, if practical, may help overcome this problem. Molecular biomarkers may assist in establishing a diagnosis[13] and identifying high risk biology,[14] but they require extensive validation before clinical use. Pathologists and clinicians also rely on metadata (age, personal and family history, lesion symptoms, recent change), which may influence diagnostic likelihoods. Importantly, it is possible to incorporate different data types, including metadata, sequential image data coupled with histopathology, to train future CNN algorithms and improve diagnostic discrimination of borderline lesions (Box 2).

Use of artificial intelligence for melanoma screening

It is well known that the incidence of invasive melanoma in Australia has increased over the past 40 years. In addition, there has been a striking increase in incidence of in situ melanoma over the past decade, from 32 cases per 100 000 population in 2004 to 80 per 100 000 population in 2019, with age-standardised mortality remaining fairly stable.[2] The potential causes for the increase in incidence are complex, and involve a true increase, driven by poor sun exposure practices of individuals born before the SunSmart era, combined with increased awareness, excessive screening, and overdiagnosis. It has recently been estimated that 54% of melanomas (15% of invasive melanomas) are overdiagnosed.[15]

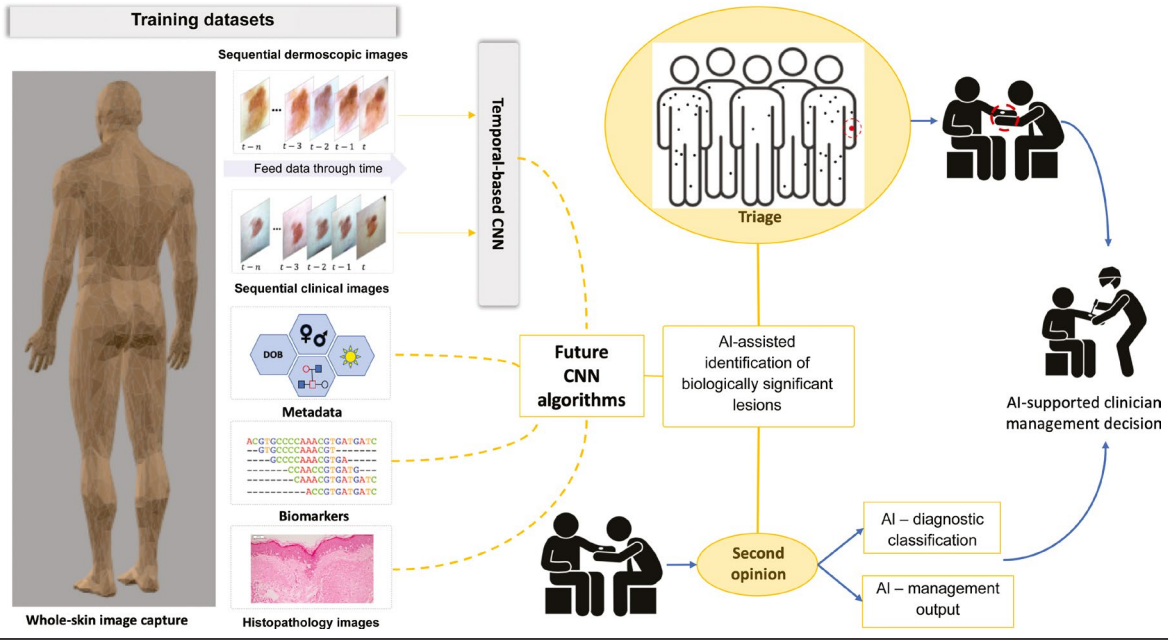
Artificial intelligence-assisted targeted screening of high risk individuals is likely to be a more effective strategy to save lives than the current opportunistic approach. With sequential whole-body image datasets linked to metadata, molecular biomarkers and clinical outcomes, our ability to identify lesions associated with sinister biological potential will improve (Box 2), thereby reducing unnecessary biopsies, minimising overdiagnosis and other potential harms associated with screening.

Use of artificial intelligence in clinical practice

There are advantages and disadvantages of introducing artificial intelligence at different points in the patient care pathway.[16] An artificial intelligence system used as a triaging tool before clinician assessment would enable automated risk stratification of individuals and/or lesions (Box 2). This approach could dramatically improve clinician workload and timely access to specialist care for people requiring urgent attention. Alternatively, artificial intelligence consulted following an examination by the clinician may act as a second opinion to improve diagnostic sensitivity and reduce unnecessary biopsies.[5] The latter is more closely aligned with current clinical workflows and therefore likely to be preferred while the field matures. There is potential for over-reliance on artificial intelligence systems in both scenarios.

A secondary support system may provide the clinician with a diagnosis or a management decision. Doctors are more likely to change their minds if they are uncertain of a diagnosis and an algorithm provides a conflicting result.[5] It is thus important to consider how an algorithm might convey uncertainty to avoid false guidance. For example, a decision-support output (eg, excise, monitor or reassure) avoids the diagnostic dilemma of differentiating between melanoma and dysplastic naevi. However, the problem is complex

2 Incorporation of different data types to train future convolutional neural network (CNN) algorithms and improve diagnostic discrimination of borderline lesions



AI = artificial intelligence. ◆

and arguments exist as to why, in many situations, a diagnostic probability output might be more desirable.

Safe implementation of new technologies

The Therapeutic Goods Administration (TGA) has developed an action plan to improve the processes by which new devices are approved for use in Australia, strengthen monitoring and follow-up, and provide more information to consumers about the devices they use.[17] International collaborations also exist with groups, such as the International Medical Device Regulators Forum, to establish better processes for medical device regulation globally. If software is classified as a medical device (ie, it is intended for diagnosis, prevention, monitoring, treatment or alleviation of disease), it must be registered on the Australian Register of Therapeutic Goods following TGA approval and before distribution within Australia.

Consumers and clinicians need to be aware of the intended use of an application or device. There are several smartphone applications available to the general public, with functionality ranging from education to monitoring and tracking to skin lesion classification. Some of these provide skin lesion risk assessment, although they may state that they are not intended to be used as a diagnostic device. There is concern that, if this is not immediately obvious to the consumer, unregistered applications may be used in lieu of seeking medical advice. Unsupervised consumer-operated diagnostic devices would require careful testing before they can be recommended.

Conclusion

As clinicians, we need to be aware of the limitations of any diagnostic tool and interpret outputs accordingly. Although the performance of artificial intelligence to date is promising, it remains to be seen how diagnostic devices in dermatology will influence decision making in the clinic and affect patient outcomes. Regardless of the specialty, any new technologies need to be rigorously tested before implementation and monitored after implementation. Ultimately, responsibility for patient care remains with the clinician and, as such, a high level of clinical acumen must be maintained. Nonetheless, artificial intelligence in dermatology is primed to become a powerful tool in skin cancer assessment.

Acknowledgements: Victoria Mar is supported by a National Health and Medical Research Council Early Career Fellowship and has an Australian Cancer Research Foundation infrastructure grant for the establishment of the Australian Centre of Excellence for Melanoma Imaging and Diagnosis, a 3D total body imaging network (Vectra, Canfield Scientific). Victoria Mar and ZongYuan Ge have received a Victorian Medical Research Acceleration Fund grant matched 1:1 by industry funding (MoleMap). We thank Scott Menzies and Yariv Levinson for sharing their knowledge in relation to regulatory requirements and processes.

Competing interests: No relevant disclosures.

Provenance: Commissioned; externally peer reviewed. ■

References are available online.

1 Mar VJ, Chamberlain AJ, Kelly JW, et al. Clinical practice guidelines for the diagnosis and management of melanoma: melanomas that lack classical clinical features. *Med J Aust* 2017; 207: 348–350. https://www.mja.com.au/journal/2017/207/8/clinical-practice-guidelines-diagnosis-and-management-melanoma-melanomas-lack

2 Australian Institute of Health and Welfare. Cancer in Australia 2019. https://www.aihw.gov.au/getmedia/8c9fcf52-0055-41a0-96d9-f81b0feb98cf/aihw-can-123.pdf.aspx?inline=true (viewed Aug 2020).

3 Erickson BJ, Korfiatis P, Akkus Z, Kline TL. Machine learning for medical imaging. *Radiographics* 2017; 37: 505–515.

4 Cabitza F, Rasoini R, Gensini GF. Unintended consequences of machine learning in medicine. *JAMA* 2017; 318: 517–518.

5 Tschandl P, Rinner C, Apalla Z, et al. Human–computer collaboration for skin cancer recognition. *Nat Med* 2020; 26: 1229–1234.

6 Rayner JE, Laino AM, Nufer KL, et al. Clinical perspective of 3D total body photography for early detection and screening of melanoma. *Front Med (Lausanne)* 2018; 5: 152.

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

8 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: 1836–1842.

9 Tschandl P, Rosendahl C, Akay BN, et al. Expert-level diagnosis of nonpigmented skin cancer by combined convolutional neural networks. *JAMA Dermatol* 2019; 155: 58–65.

10 Hekler A, Utikal JS, Enk AH, et al. Deep learning outperformed 11 pathologists in the classification of histopathological melanoma images. *Eur J Cancer* 2019; 118: 91–96.

11 Fujisawa Y, Otomo Y, Ogata Y, et al. Deep-learning-based, computer-aided classifier developed with a small dataset of clinical images surpasses board-certified dermatologists in skin tumour diagnosis. *Br J Dermatol* 2019; 180: 373–381.

12 Elmore JG, Barnhill RL, Elder DE, et al. Pathologists' diagnosis of invasive melanoma and melanocytic proliferations: observer accuracy and reproducibility study. *BMJ* 2017; 357: j2813.

13 Clarke LE, Warf MB, Flake DD, et al. Clinical validation of a gene expression signature that differentiates benign nevi from malignant melanoma. *J Cutan Pathol* 2015; 42: 244–252.

14 Tang DY, Ellis RA, Lovat PE. Prognostic impact of autophagy biomarkers for cutaneous melanoma. *Front Oncol* 2016; 6: 236.

15 Glasziou PP, Jones MA, Pathirana T, et al. Estimating the magnitude of cancer overdiagnosis in Australia. *Med J Aust* 2019; 212: 163–168. https://www.mja.com.au/journal/2020/212/4/estimating-magnitude-cancer-overdiagnosis-australia

16 Janda M, Soyer HP. Can clinical decision making be enhanced by artificial intelligence? *Br J Dermatol* 2019; 180: 247–248.

17 Therapeutic Goods Administration. Medical devices reforms. https://www.tga.gov.au/medical-devices-reforms (viewed Aug 2020). ■