Machine learning in clinical practice: prospects and pitfalls

Machine learning has huge potential to enhance clinical decision making, but there are still many limitations

Machine learning (ML), a subdiscipline of artificial intelligence, encompasses a family of computerised (machine) methods that identify (learn) patterns in large (training) datasets not detectable to humans (Box 1). Identified patterns are then encoded in a computer model or algorithm which is then tested and validated on new data. Three basic ML types exist (Box 2), with supervised and reinforcement learning being used most frequently.

Algorithms can take various forms — deep neural networks (deep learning systems) currently dominate.[1] These networks comprise algorithms that cluster and classify information in a manner resembling the human brain. Just as neural synapses are strengthened through repeated activity, deep neural networks iteratively strengthen their functions through mathematical means, adjusting the weights of inputs as they move through layers of intermediate nodes (neurons) towards a desired output.

Recent renewed interest in ML is driven by the availability of massive digitised datasets from genomics, biobanks, medical images, administrative datasets, electronic health records and wearable biosensors; advances in computer processing speed, storage and internet-mediated computing power; and increasing commercial investment in ML.[2] Here, we introduce ML applications with potential clinical utility while considering current limitations. More exhaustive reviews of ML methodology and health care uses can be found elsewhere.[3,4]

Emerging machine learning applications

Clinical applications of artificial intelligence systems built using ML are being used to diagnose conditions, estimate risk, select optimal treatments and determine drug dosing. Some applications can perform as well as, if not faster or more accurately than, clinicians, and do so more consistently, devoid of human error from fatigue. ML applications are currently most advanced for diagnoses based on imaging data and risk prediction.

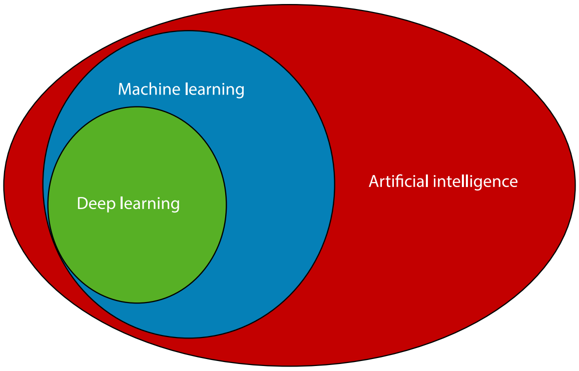
Diagnostic applications

Retinal cameras incorporating ML algorithms can confer high diagnostic accuracy in general practice in the early diagnosis of referable diabetic retinopathy, possible glaucoma and age-related macular degeneration. In one study, a deep learning system was trained on a total of 494 661 retinal images to detect these conditions in community and clinic-based multi-ethnic populations with diabetes. The deep learning system was then tested on 71 896 images from 14 880 patients, and its results were validated by trained eye specialists. The deep learning system diagnosed referable diabetic retinopathy, possible glaucoma and age-related macular degeneration with sensitivity/specificity of 90.5%/91.6%, 96.4%/87.2%, and 93.2%/88.7%, respectively.[5] Similarly, another deep learning system demonstrated sensitivity, specificity and discrimination (area under the receiver operator characteristic curve [AUC]) for diagnosing referable diabetic retinopathy of 87.0–90.3%, 98.1–98.5% and 0.990–0.991 (0.80–1.00 representing good to perfect discrimination), respectively.[6] A successful trial in primary care led the United States Food and Drug Administration to list a commercial retinal camera system as its first ML-based medical application.[7]

In dermatology, early and accurate diagnosis of suspicious skin lesions could be aided by smartphone cameras with ML-based apps, allowing individuals to conveniently perform their own regular skin checks and transmit results to specialists who may not be easily accessible. Trained on 129 450 dermoscopy images of 2032 different skin conditions, a deep learning system was tested against 21 dermatologists using 1700 biopsy-proven examples of keratinocyte carcinomas, benign seborrhoeic keratoses, melanomas and benign nevi as reference.[8] The deep learning system in all cases matched and, in some cases, out-diagnosed the dermatologists.

Accurate diagnosis of breast cancer metastases in axillary lymph node biopsies is vital in determining

1 Relationship between artificial intelligence, machine learning and deep learning



Artificial intelligence describes any computer method that mimics human reasoning capabilities, including pattern recognition, abstract reasoning and planning. Machine learning is one of several subdomains of artificial intelligence and comprises a family of methods including neural networks, information theory and probabilistic approaches to learn new knowledge from past examples. Deep learning is one of many approaches to machine learning and involves training multilayered neural networks on large datasets to learn patterns to perform tasks such as speech and image recognition. ◆

Ian A Scott[1,2]
David Cook[1]
Enrico W Coiera[3]
Brent Richards[4]

1 Princess Alexandra Hospital, Brisbane, QLD
2 University of Queensland, Brisbane, QLD.
3 Centre for Health Informatics, Macquarie University, Sydney, NSW.
4 Gold Coast Hospital and Health Service, Southport, QLD.

ian.scott@health.qld.gov.au

doi: 10.5694/mja2.50294

2 Different types of machine learning

Type	Uses	Example
Supervised learning: maps input data to known outputs	Classification: distinguishing between different items, categories or subgroups Prediction (or regression): predicting values of an output variable based on input variables	Making a diagnosis or predicting risk of a clinical event based on provided risk factors or laboratory results
Unsupervised learning: only input data are provided and the model must identify or learn relationships without reference to known outputs	Clustering: identifying clusters that appear to share latent similarities and recognising cluster features Anomaly detection: recognising unusual patterns in values for different variables within datasets	In patients with the same diagnosis but different responses to therapy, identifying features such as genomic or phenotypic profiles that predict response to treatment
Reinforcement learning: models learn an optimised set of rules for achieving a goal or maximising an expected return by a process of trial and error	Useful when the system is dynamic and the model must adapt to change, or the basic function is known but automated tuning of predictions or actions is desirable	Choosing ventilator and vasopressor settings in patients with severe sepsis in intensive care units

treatment. In a simulation challenge, 32 ML algorithms were trained on data from whole-slide images with and without metastases, as determined by immunohistochemical staining.[9] When applied to an independent test set of 129 whole-slide images (49 with and 80 without metastases) and performance compared with that of 11 pathologists examining the same slides under a time constraint-simulating routine pathology workflow, the best performing algorithm outperformed the pathologists (AUC, 0.994 *v* mean, 0.810 [range, 0.738–0.884]).[9]

Echocardiography requires considerable expert operator time in measurement and interpretation, limiting its use in primary care and rural settings — limitations which ML may overcome. An automated deep learning system was trained on 14 035 echocardiographs to quantify cardiac chamber volumes, left ventricular mass and ejection fraction, and detect hypertrophic cardiomyopathy, cardiac amyloid and pulmonary arterial hypertension.[10] When applied to 8666 echocardiograms performed by ultrasonographers during routine workflows using commercial software, automated measurements were comparable or superior across 11 internal consistency metrics, and hypertrophic cardiomyopathy, cardiac amyloid and pulmonary arterial hypertension were identified with AUCs of 0.93, 0.87 and 0.85, respectively.[10]

Risk prediction

Current prediction tools may lack generalisability by virtue of a limited set of preselected variables judged to be clinically relevant. In contrast, ML can utilise many more variables available through electronic health records and may better predict patient trajectories across diverse populations. These outputs could inform models of care, resource allocation and targeting of care to high risk patients.

Electronic health record data from 194 470 admissions across two US hospitals were used to train ML algorithms to predict in-hospital mortality, long length of stay, discharge diagnoses 24 hours after admission, and 30-day unplanned readmission risk at discharge.[11] When tested on another 21 751 admissions across both hospitals and compared with traditional prediction tools, algorithms more accurately predicted in-hospital mortality (AUC, 0.93–0.95 *v* 0.85–0.86), long length of stay (AUC, 0.85–0.86 *v* 0.74–0.77) and readmission risk (AUC, 0.76–0.77 *v* 0.68–0.70), and classified all discharge diagnosis codes with weighted AUCs of 0.86 and 0.87.[11]

Regarding specific conditions, identifying high risk septic patients may guide resuscitation and treatment efforts. In a retrospective study, an ML algorithm trained on 4222 admissions using over 500 clinical variables was compared with five commonly used regression-based prediction tools in a test set of 1056 admissions. The algorithm outperformed all other tools with an AUC of 0.86 versus a range of 0.69–0.76.[12] Predicting in-hospital survival among patients with major head trauma can help decide indications for neurosurgery and rehabilitation. Using data from 7769 patients with computed tomography head scans showing brain damage, five ML algorithms trained on 11 input variables had performance on 100 novel patients compared with that of four traditional prediction models and the opinions of ten neurosurgeons. Algorithms were more accurate, more sensitive, as specific and more discriminating than both traditional models and clinicians, with a mean AUC of 0.86 versus 0.77 and 0.74, respectively.[13]

Other ML applications are in development (Box 3).[14–18] Australian researchers are actively engaging in ML,[19,20] with the potential to do so at scale because every citizen has a unique Medicare identifier, and large digitised data repositories are evolving with state-wide digital hospital systems and My Health Records, all protected by strong privacy legislation.

Limitations and challenges

Dependence on data quality

Data that are incomplete, incorrect (including wrong diagnoses), poorly described or labelled, inadequately structured (semantically or temporally), obsolescent, unrepresentative of diseases or populations of interest, or of low volume will introduce error in ML training. Errors can reflect random omissions or misclassifications, or more importantly, systematic

3 Applications of machine learning in development

Domain	Examples
Predicting response to treatment	Predicting the most effective anti-HIV treatment for any patient and virus combination; machine learning algorithms compare favourably with the most commonly used genotype interpretation systems and HIV drug resistance expertise[14] Predicting optimal use of intravenous fluids and vasopressors for patients admitted with severe sepsis[15]
Improving efficiency in various clinical domains	Determining optimal dosing regimens for various medications[16] Selecting eligible patients for faster enrolment in clinical trials[17] Accelerating drug discovery by ranking associations of biomarkers with different diseases[18]

HIV = human immunodeficiency virus. ◆

biases in data collection, for example, regarding race, ethnicity, language, socio-economic status and sexual preference.

Poorly constructed algorithms can hinder decision making

Inner workings of ML algorithms can be opaque and uninterpretable to clinicians, and poorly constructed or improperly used algorithms can impair decision making. For example, an ML model predicting survival of post-menopausal women performed worse than conventional Framingham scores, partly because it lacked training information on key blood investigations.[21] Another ML model falsely classified asthmatic patients with pneumonia as being low risk and eligible for early hospital discharge, because training data did not capture life-saving admissions of many patients to intensive care units.[22] Algorithms have been developed to predict optimal treatments for various cancers, but after 4 years, none have demonstrated superior performance to oncologists, with systems still struggling with simply identifying cancer types.[23]

Data relating to clinical features irrelevant to outcomes of interest (eg, eye colour and cholesterol levels in predicting antibiotic response in appendicitis) can render algorithms excessively complex and diminish predictive accuracy. Clinical experts working alongside data experts must identify the most pertinent features.

Need for reference standards

For algorithms developed using supervised learning, a robust reference standard must be included in both training and testing datasets for each diagnosis (gold standard testing or assessment protocols) or outcome (appropriately specified, verifiable event or outcome criteria).

Insensitivity to context and timing of events

In defining predictor variables or associations between variables, algorithms may be insensitive to contextual factors, such as local clinician preferences, care standards or admission policies. Temporal variations in variables or sequences of clinical decisions are other confounders. For example, severely ill septic patients may appropriately receive fluids earlier than healthier patients, yet be more likely to die. ML correctly associates earlier fluid administration with higher mortality, which may be misinterpreted as contributing to death. Similarly, ML can perpetuate prior patterns of poor decisions and errors recorded in electronic health records or coded data on which ML is trained. These examples underscore the need for expert clinician interpretation of ML outputs.

Lack of defined evidence standards

Standards for assessing safety and utility of ML applications are currently not well defined. Recent United Kingdom guidance[24] recommends randomised trials for ML algorithms that directly affect patient care, and these are starting to appear.[25] Replication studies which test validity, reproducibility and generalisability are also needed.[26]

Lack of impact studies

The value of ML is gauged by increased diagnostic accuracy and therapeutic effectiveness, decreased time on routine tasks, faster turnaround of investigation results, reduced costs of care, and better patient outcomes. Clinical impact studies and cost–benefit analyses of ML in routine care are mostly lacking. Implementing ML algorithms in routine clinical workflows requires making interfaces accessible within clinical information systems, avoiding both alert fatigue and unquestioning acceptance of ML predictions, determining liability if patient harm ensues, and ensuring privacy of patient and practitioner data.

Conclusion

While ML will likely disrupt clinical practice over coming decades, particularly imaging-based disciplines, it requires judicious application. ML can provide better, more patient-specific information, affording clinicians greater capacity to make the most appropriate clinical decisions in partnership with their patients.

Competing interests: Brent Richards has received non-financial support from Amazon Web Services and non-financial support from Microsoft. ■

References are available online.

1 Esteva A, Robicquet A, Ramsundar B, et al. A guide to deep learning in healthcare. *Nature Med* 2019; 25: 24–29.

2 Naylor CD. On the prospects for a (deep) learning health care system. *JAMA* 2018; 320: 1099–1100.

3 Rajkomar A, Dean J, Kohane I. Machine learning in medicine. *N Engl J Med* 2019; 380: 1347–1358.

4 Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nature Med* 2019; 25: 44–56.

5 Ting DSW, Cheung CY-L, Lim G, 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: 2211–2223.

6 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: 2402–2410.

7 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: 39.

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

9 Ehteshami Bejnordi B, Veta M, van Diest PJ, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA* 2017; 318: 2199–2210.

10 Zhang J, Gajjala S, Agrawal P, et al. Fully automated echocardiogram interpretation in clinical practice. Feasibility and diagnostic accuracy. *Circulation* 2018; 138: 1623–1635.

11 Rajkomar A, Oren E, Chen K, et al. Scalable and accurate deep learning with electronic health records. *npj Digit Med* 2018; 1: 18.

12 Taylor RA, Pare JR, Venkatesh AK, et al. Prediction of in-hospital mortality in emergency department patients with sepsis: a local big data-driven, machine learning approach. *Acad Emerg Med* 2016; 23: 269–278.

13 Rughani AI, Dumont TM, Lu Z, et al. Use of an artificial neural network to predict head injury outcome: clinical article. *J Neurosurg* 2010; 113: 585–590.

14 Zazzi M, Incardona F, Rosen-Zvi M, et al. Predicting response antiretroviral treatment by machine learning: the EuResist project. *Intervirology* 2012; 55: 123–127.

15 Komorowski M, Celi LA, Badawi O, et al. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nature Med* 2018; 24: 1716–1720.

16 Nemati S, Ghassemi MM, Clifford GD. Optimal medication dosing from suboptimal clinical examples: a deep reinforcement learning approach. *Conf Proc IEEE Eng Med Biol Soc* 2016; 2016: 2978–2981.

17 Ni Y, Kennebeck S, Dexheimer JW, et al. Automated clinical trial eligibility prescreening: increasing the efficiency of patient identification for clinical trials in the emergency department. *J Am Med Inform Assoc* 2015; 22: 166–178.

18 Stephenson N, Shane E, Chase J, et al. Survey of machine learning techniques in drug discovery. *Curr Drug Metab* 2019; 20: 185–193.

19 Maali Y, Perez-Concha O, Coiera E, et al. Predicting 7-day, 30-day and 60-day all-cause unplanned readmission: a case study of a Sydney hospital. *BMC Med Inform Decis Mak* 2018; 18: 1.

20 Nanayakkara S, Fogarty S, Tremeer M, et al. Characterising risk of in-hospital mortality following cardiac arrest using machine learning: a retrospective international registry study. *PLoS Med* 2018; 15: e1002709.

21 Gorodeski EZ, Ishwaran H, Kogalur UR, et al. Use of hundreds of electrocardiographic biomarkers for prediction of mortality in postmenopausal women: the Women's Health Initiative. *Circ Cardiovasc Qual Outcomes* 2011; 4: 521–532.

22 Caruana R, Lou Y, Gehrke J, et al. Intelligible models for healthcare: predicting pneumonia risk and hospital 30-day readmission. Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2015 Aug 10-13; Sydney, Australia. https://www.microsoft.com/en-us/research/wp-content/uploads/2017/06/KDD2015FinalDraftIntelligibleModels4HealthCare_igt143e-caruanaA.pdf (viewed Nov 2018).

23 Ross C, Swetlitz I. IBM pitched its Watson supercomputer as a revolution in cancer care. It's nowhere close. STAT Investigation. 5 Sept 2017. https://www.statnews.com/2017/09/05/watson-ibm-cancer/ (viewed Nov 2018).

24 National Institute for Heath and Care Excellence. Evidence standards framework for digital health technologies. London: NICE, 2019. https://www.nice.org.uk/Media/Default/About/what-we-do/our-programmes/evidence-standards-framework/digital-evidence-standards-framework.pdf (viewed Mar 2019).

25 Shimabukuro DW, Barton CW, Feldman MD, et al. Effect of a machine learning-based severe sepsis prediction algorithm on patient survival and hospital length of stay: a randomised clinical trial. *BMJ Open Resp Res* 2017; 4: e000234.

26 Coiera E, Ammenwerth E, Georgiou A, Magrabi F. Does health informatics have a replication crisis? *J Am Med Inform Assoc* 2018; 25: 963–968. ■