JIM Review

doi: 10.1111/joim.12822

eDoctor: machine learning and the future of medicine

G. S. Handelman[1], H. K. Kok[2], R. V. Chandra[3,4], A. H. Razavi[5,6], M. J. Lee[7] & H. Asadi[3,8,9]

From the [1]Royal Victoria Hospital, Belfast, UK; [2]Interventional Radiology Service, Northern Hospital Radiology, Epping; [3]Interventional Neuroradiology Service, Monash Imaging, Monash Health; [4]Faculty of Medicine, Nursing and Health Sciences, Monash University, Clayton, Vic, Australia; [5]School of Information Technology and Engineering, University of Ottawa; [6]BCE Corporate Security, Ottawa, ON, Canada; [7]Department of Radiology, Beaumont Hospital and Royal College of Surgeons in Ireland, Dublin, Ireland; [8]Department of Radiology, Interventional Neuroradiology Service, Austin Health, Heidelberg; and [9]School of Medicine, Faculty of Health, Deakin University, Waurn Ponds, Vic, Australia

Abstract. Handelman GS, Kok HK, Chandra RV, Razavi AH, Lee MJ, Asadi H (Royal Victoria Hospital, Belfast, UK; Northern Hospital Radiology, Epping; Monash Imaging, Monash Health; Monash University, Clayton, Vic, Australia; University of Ottawa; BCE Corporate Security, Ottawa, ON, Canada; Austin Health, Heidelberg, Vic, Australia; Beaumont Hospital and Royal College of Surgeons in Ireland, Dublin, Ireland; Deakin University, Waurn Ponds, Vic, Australia). eDoctor: machine learning and the future of medicine. (Review) *J Intern Med* 2018; **284**: 603–619.

Machine learning (ML) is a burgeoning field of medicine with huge resources being applied to fuse computer science and statistics to medical problems. Proponents of ML extol its ability to deal with large, complex and disparate data, often found within medicine and feel that ML is the future for biomedical research, personalized medicine, computer-aided diagnosis to significantly advance global health care. However, the concepts of ML are unfamiliar to many medical professionals and there is untapped potential in the use of ML as a research tool. In this article, we provide an overview of the theory behind ML, explore the common ML algorithms used in medicine including their pitfalls and discuss the potential future of ML in medicine.

Keywords: artificial intelligence, machine learning, medicine, supervised machine learning, unsupervised machine learning.

Introduction

Artificial intelligence (AI) has promised to revolutionize medicine for over 30 years, and there have been technological breakthroughs in recent years that could make this a reality, including exponential increases in computing power, big-data processing technologies, access to large clinical data sets using electronic health records, and machine learning (ML) [1]. ML in medicine can lead to more accurate diagnostic algorithms and individualize patient treatment [2, 3]. To help clinicians develop a better understanding of ML, we will review ML as it applies to medicine in two areas; first by looking at the core concepts and algorithms of ML applicable to medicine and following this, we will review the current use of ML by various specialties and illustrate how commercial interests have started to expand into this area of medicine and the potential for an 'eDoctor' on the horizon [4].

Machine Learning (ML) at a glance

The name machine learning (ML) was initially coined in 1959 by Arthur Samuel, a prominent computer scientist at the time. The terms AI and ML are often used interchangeably, though incorrectly; AI refers to the overarching concept of the 'thinking machine' or automated decision-making whereas he described ML as giving 'computers the ability to learn without being explicitly programmed' [5, 6]. The main premise of ML is to introduce algorithms that ingest input data, apply computer analysis to predict output values within an acceptable range of accuracy, identify patterns and trends within the data and finally learn from previous experience. ML is not new – since the inception of modern computing, this idea of the thinking machine has been proposed with the aim of applying the computational capacity of computers towards elucidating patterns and conclusions that would be difficult to reach by conventional statistical methods which rely on human operators

to devise and supply a rule base or assumption on correlation to the computer for further analysis [7].

With ML, the process is semiautomated – the computer is provided the data and it creates complex analytical models based on a learning framework to refine and optimize the accuracy of prediction. That is not to say it is inherently divergent to conventional statistics as in many ways ML is either based on or adopts statistical underpinnings to how it works [8]. Early iterations of the concept date back to the 1950s with Turing's 'learning machine' and the development of the first neural network focusing mainly on military experiments. Over the next few decades, stepwise academic improvement and discovery led to initial commercialization in the 1980s, with early work focused on the life sciences with concomitant collaboration of computer scientists and medical scientists resulting in the creation of computing resources such as the Advanced Research Projects Agency Network (ARPANET) and the National Science Foundation Network (NSFNET), allowing researchers to address medical problems with AI methods [9]. In 1985, the Artificial Intelligence in Medicine (AIME) conference was inaugurated, focusing on those elements that linked computer science, medicine and biology because of increasing awareness that the computational power of computers could be clinically useful and a journal ('Artificial Intelligence in Medicine') dedicated to research within the field commenced in 1989 [9]. Over the ensuing years, presentations at AIME have evolved from knowledge engineering to increasingly data-driven techniques, ML, bioinformatics and semantic technology with much of this change brought about by the influence of statistics and probability theory. Once data storage became larger, cheaper and connected, the integration of data mining and large data analytics into the paradigm increased the scope and potential of ML and was similarly reflected in the theme of the conference [10]. More recently, many existing conferences are often adding presentations on different aspects of ML and there has been an expansion of conferences dedicated to ML in health care; 'Artificial Intelligence in Medicine (AIMed)' which has separate conferences in Europe, North America and Asia, the 'Human Intelligence and Artificial Intelligence in Medicine' symposium and 'Machine Learning in Healthcare' meeting in Stanford University, the 'Deep Learning in Healthcare Summit' in Boston, 'Machine Learning, Big Data and AI in Healthcare' conference in Washington and the 'Predictive Analytics World for Healthcare' in Las Vegas.

Conventional statistics is largely based on the testing of a hypothesis around cause and effect and chooses models around significance and in-sample goodness of fit. ML is less focused on the interpretability of models and mathematically focuses on the predictive performance and generalization of models around cross-validation and iterative improvement of the algorithm. It is frequently associated with versatility within high-dimensional space which has a large variety of variables in differing formats, utilizing feature selection (selecting subsets of relevant features for use in model creation), pattern analysis and dimensionality reduction (reducing the number of random variables used by reducing data to its principal components) for complicated problems such as gene expression analysis, text analysis and computational chemistry [11, 12].

A basic explanation of an ML process can be achieved by considering the example of training a computer to detect cancer from histopathological slides. One could, by using an annotated knowledge base, try to program the computer to use a rule base to detect the combinations of colours and lines that represent invasion of disease through a basement membrane and then when presented with a new slide, the computer program could provide a confidence score or likelihood of malignancy. Alternatively, a database of images consisting of malignant and nonmalignant specimens could be presented to allow the computer program to determine how to best differentiate between the two categories and then judge the success of the program by presenting it with new slides to verify its accuracy. If successful, the process by which the computer reaches its conclusions is of less importance if it is better at predicting the final result than our current knowledge and rule-based system [13].

More recently, the term 'deep learning' has entered popular lexicon through attention in the media and is sometimes equated with AI. However, as we detail below, deep learning is a tool that is part of the continuum of the various statistical and probabilistic algorithms available. The mathematics and methodological details of ML are beyond the scope of this article and for most readers are not essential. However, a few important concepts will be elaborated upon.

Primary amongst these is supervised and non-supervised learning (Fig. 1).

Supervised learning

In supervised learning, the computer is provided with features related to the learning target (such as patient demographics and risk factors) and desired outcome measures to be achieved (such as diagnoses or clinical events) with the goal of identifying links between those two in the data set. A commonly used example is training a model to differentiate between apples, oranges and lemons [14]. The 'label' of each type of fruit is initially supplied to the algorithm along with the features such as colour, size, weight and shape and the algorithm learns the mix of features that differentiate the fruits. Then, when a new, 'unlabelled' fruit is presented, the model should be able to predict which type of fruit it is.

This approach has been used in histological specimens and has even been crudely used in 'reverse image searching' by using images of specimens and utilizing the search engine to find similar images [13, 15]. Similarly, this process of deriving output variables from previous examples of known input variables allows for regression analysis, a concept commonly used in statistics with the difference being potential iterative improvement in accuracy of prediction by learning to progressively augment the prediction algorithm along with ability to deal with more variables and model complex nonlinear relationships between independent and dependent

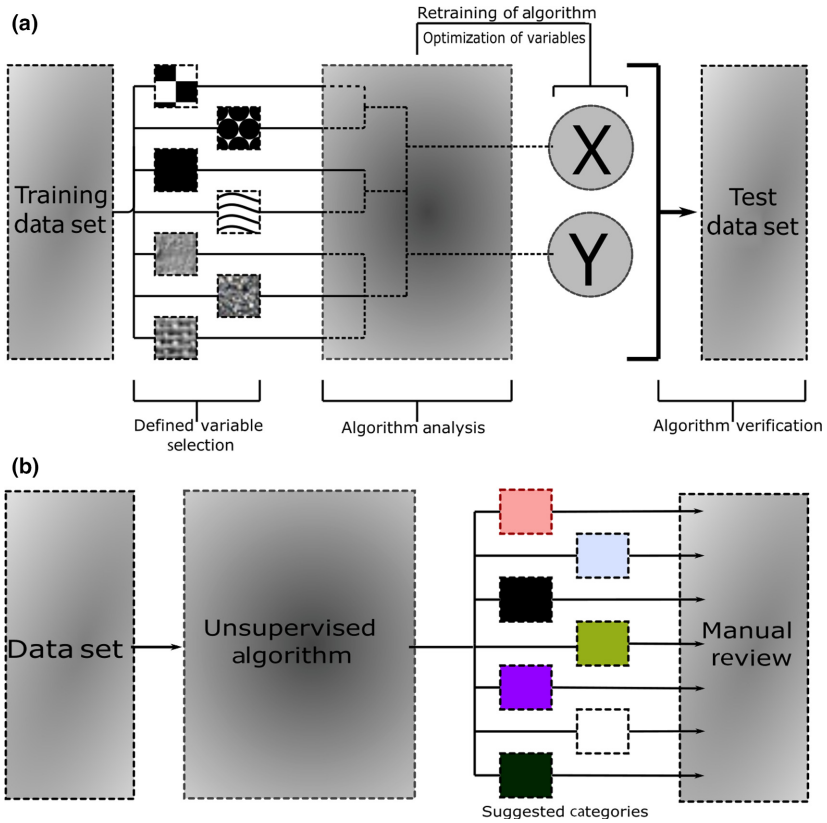

Fig. 1 *(a) Flow diagram of a supervised ML. A training data set containing representative data (e.g. demographic information or risk factors) as well as outcomes of interest (e.g. mortality) is analysed by the ML algorithm to identify patterns or links in the data presented for training. The system has the ability to continuously retrain under supervision to improve its accuracy. When it has been optimized with the training data set, the algorithm can then be applied to a 'true' test data set of interest. (b) Flow diagram of unsupervised ML. A large data set is presented to the unsupervised algorithm where patterns are elucidated and presented to the user for manual review.*

variables [16]. Thus, supervised learning techniques can be thought to be primarily focused on *classification* – to identify categories (subpopulations) of a new observation, based on a training set of data comprising observations (or instances) whose category membership is known and *regression* – predicting continuous values for a target variable based on training set instances with known values of that variable such as when calculating risk of cardiovascular disease, predicting tumour size, estimating individualized disease-free survival or predicting length of hospital stay [17].

Once the model has been created and optimized, it is tested on novel patients not included in the training data to determine its external validity and thus, applicability to other patients. In medicine, it is most commonly applied to clinical problems related to diagnosis and prognosis. For diagnosis, training the algorithm is done using a training data set and having disease presence or absence as desired outcome variables and then validating the model on a separate test data set. Similarly, outcome or prognosis can be predicted by using a training data set and (for example) setting 5-year survival or disease-free survival as desired outcome measures and validating the trained model on a separate data set [18]. Conventional statistical techniques such as Cox proportional hazard, logistic regression and ridge regression can be integrated into the ML paradigm whilst some of the nomenclatures of ML are synonyms for functions in conventional statistics due to the converging approaches of computer scientists and statisticians into a single speciality [8, 19, 20].

Unsupervised learning

In unsupervised learning, the computer is provided with unclassified data records to recognize and determine whether any existing latent patterns are present, sometimes producing both answers and questions that may not have been conceived by the investigators. From a technical standpoint, whilst supervised learning primarily deals with *classification* and *regression* problems, unsupervised learning deals more with *clustering* and *dimensionality reduction*. The patterns identified in unsupervised learning commonly have to be evaluated for utility either by human interrogation or via application within a supervised learning task.

Clustering refers to the identification of groups within data, that is, the algorithm is provided the data, analyses it and determines any latent similarities within data that allows subjects to be grouped into subsections and patterns within the data to be discovered (Fig. 2). Due to the complexity of data and heterogeneity of patients in medicine, identifying groups can be difficult by intuition and the ability to elucidate these groups allows more targeted diagnostics, therapeutics and prognostication. As an example, Vranas *et al.* looked at the data heavy setting of the intensive care unit where resource burden and outcome prediction are not inconsiderable problems. By retrospectively analysing patient variables using an unsupervised algorithm (clustering), they were able to identify patient subgroups that had significantly different clinical courses despite similar diagnoses and when applying the predictive clusters to a separate, unseen data set, the predictive capability persisted [21].

A common application of such a process is to explore complex interrelationships between genetics, biochemistry, histology and disease states within the data. Methods such as logistic regression, principal component analysis and latent class analysis have been used for many years; however, it is the integration with iterative supervised and unsupervised learning techniques that novel applications have emerged [13, 14, 22, 23].

Anomaly detection, the process by which an ML algorithm interrogates data to look for unusual patterns, is a branch of ML that has had some interest. It is useful for finding patients that do not fit into the average clinical trajectory and perhaps require further investigation, alteration in treatment or consideration of medical error [24]. In this setting, the algorithm is trained on a set of patient cases where the diagnoses or outcomes are known and when applied to new patient cases, alerts the doctor when one of these does not seem to fit into the standard patient pattern. As an example, Hauskrecht *et al.* [25] took a database of patients diagnosed with pneumonia along with their clinical and biochemical characteristics on presentation and designed an anomaly detection algorithm that would detect when there was a decision to treat (at home or admit to hospital) which deviated from the norm. Similar application of this type of methodology has been applied to detecting abnormal electroencephalography (EEG) data [26], and even detection of outliers so that they can be interrogated to see whether they are truly representative

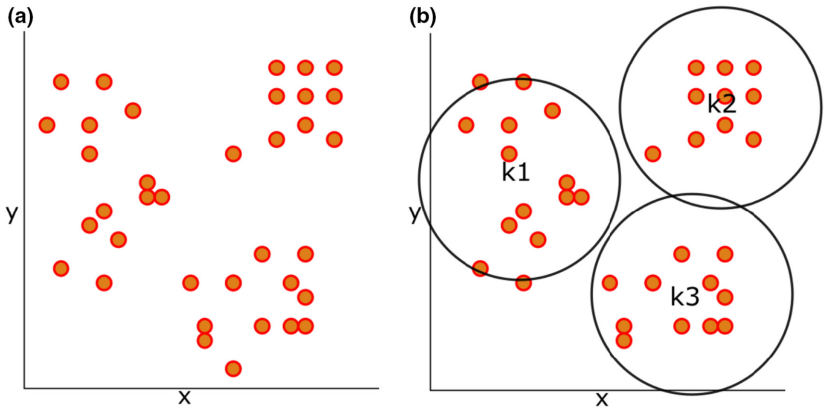

Fig. 2 *Example of a clustering algorithm, 'k-means'. The algorithm is asked to cluster data into* k *number of groups and does so based on similarity measures.*

or should be removed from the data set to improve model accuracy [27].

Semisupervised learning is an amalgamation of supervised and unsupervised ML that can analyse a large amount of unlabelled (unclassified) data whilst augmenting its pattern recognition abilities with a small amount of labelled (preclassified) data. From a medical perspective, this approach is valuable, since assigning labels to information (e.g. patient records) can be time-consuming and costly, given the complexity and abundance of medical data [28]. Moreover, semisupervised learning can increase the speed and accuracy of information extraction from large data sets. Semisupervised learning has been used to analyse scientific articles for inclusion in systematic reviews of a topic to decrease the amount of work involved in detailed literature reviews, predict progression of dementia and for the detection of breast cancer from images [29–31].

ML and medicine

Machine learning has grown in recent years due to advances in computer science, spurred on by the technology industry, which relies heavily on ML for a variety of applications [26, 27]. Medicine has not been immune to this progress and is in fact, a fertile ground for ML. As discussed previously, the concept of applying AI to medicine is not new, with applications dating back to the 1970s but many medical professionals are not familiar with ML as a concept, how it could be applied or the breadth of publications on ML which already exists within their own specialties.

The work already completed suggests the potential for improving prediction and visualization quality in research if ML is adopted and applied. With continued improvements in ML techniques in medicine, researchers and clinicians could face being left behind as paradigms in healthcare change. To better illustrate this potential, we present a selection of studies and applications of ML across a range of medical specialties.

Personalized medicine

ML for personalized medicine is a growing area of interest [32]. The ability to draw on large data sets and predictive models allows for clinicians to more confidently diagnose, predict and treat their patients. Personalized medicine is individualization, recognition of the microvariables within a patient that may cause them to be phenotypically different to their peers. The power of ML allows for this individualization at many levels: diagnosis [33], prognosis and treatment [34, 35]. Within this framework, it is perhaps difficult for the clinician to be able to predict ahead of time what clusters of patients exist and this is where unsupervised learning lends power to the data processing; it does not require 'solved' cases or predetermined classes to be provided in order to make associations and predictions from the data, and studies have shown that previously unidentified patient subgroups can be determined by interrogation of data by an unsupervised method [36, 37].

Salgado *et al.* [38] described the use of ML to predict the need for vasopressor administration based on 24 clinical variables commonly recorded in the intensive care setting by using unsupervised learning to extract features to use in modelling and apply it to individual cases with reasonable success. Weng *et al.* used routine clinical data on 378 256 patients from the UK Clinical Practice Research Datalink (CPRD) used by general practitioners to train a variety of ML algorithms to estimate risk of cardiovascular events and compared this to the current American College of Cardiology/American Heart Association (ACC/AHA) guideline's knowledge-based algorithm and found that not only was the algorithm superior, it also identified significant clinical variables that were not included in other predictive scores, such as chronic obstructive pulmonary disease and severe mental illness [39]. Databases such as these are invaluable and will in future allow creation of population and patient-specific decision support tools.

Therapeutics

Drug discovery and pharmacokinetic prediction have been enhanced with ML and deep learning techniques [40], a fact that Gawehn *et al.* [41] expanded upon in their review of deep learning in drug discovery, claiming that the addition of this tool is 'game-changing' and will be useful in toxicity prediction, genome mining and chemogenetic applications.

Focusing on drug discovery alone would merit its own review article and has been covered previously by Kirchmair *et al.*'s article in *Nature* [40]. However, aside from pharmaceutical research, there are novel ways in which ML has been applied to therapeutics. In an article by Stanfield *et al.* [42], neural networks were applied to a combination of cell line mutation data, previously known response rates and protein–protein interactions to predict response rates to oncology drugs with an accuracy of 85%. Again, their integration of data from the Genomics of Drug Sensitivity in Cancer (GDSC) and the Cancer Cell Line Encyclopedia (CCLE) projects underscores the importance of such databases and bioinformatics data to personalized medicine.

Surgery

Within the field of surgery, ML has been applied most commonly to two main areas: robotics and decision support. Kassahun *et al.* [43] provide a succinct overview of ongoing research and advances in surgical robotics including autonomous endoscopic guidance or autonomous knot tying, executing simple tasks with performance exceeding human capabilities, as well as discovering new operating techniques that are superior to current practice. Nomograms and decision aides have been presented for years, and many clinicians will be familiar with these and commonly use the online iterations of these [44]. ML decision support systems, similar to medicine and diagnostics, are becoming common in identifying surgical candidates, potential postoperative complications, making diagnoses and delivering outcome predictions [33,34].

As an example of this, Kiranantawat *et al.* [45] described using smartphone images of iatrogenically induced tissue ischaemia to build a predictive algorithm using k-nearest neighbour (a ML technique of pattern recognition and regression) that could, with a sensitivity of 94% and specificity of 98%, detect venous or arterial occlusion and was also able to distinguish between venous, arterial or mixed vascular occlusion by taking smartphone pictures of subject's fingers.

Radiology

ML techniques are growing in radiology, however, as a speciality that deals almost exclusively with data in the form of predominantly digital requests, images and linked text reports, it is a speciality that is ripe for ML [22]. The potential for pattern recognition and automated analyses of images in radiology is unique and is perhaps why this area is beginning to be commercially developed with some arguing that automation will replace much of the radiologists' work [46].

Due to the increase in number of imaging modalities, resolution and number of images generated from modern scanners, computer-aided image analysis is becoming more prevalent in reducing the radiologists' workload. ML has been applied to medical image segmentation, whereby the computer learns to separate structures and organs [47–49], as well as registration, computer-aided detection (CAD) [50–52] and brain functional analyses [41,42]. Content-based image retrieval systems have also been proposed and designed, whereby a database of images can be searched based on the similarities between images [53, 54],

empowered by text analysis of radiology reports [22].

Summers *et al.* [51] reported on ML CAD of polyps on CT colonography images, where the authors took 1186 patients with CT colonography images verified by optical colonoscopy, divided the data into training and test data sets and trained a polyp CAD program on the data, showing no significant difference in sensitivity between ML and optical colonoscopy for detecting polyps larger than 8 mm. In the past, similar attempts have been made to detect lung nodules and pulmonary emboli and commercial CAD programs are already in use to increase the sensitivity of screening; however, concerns have been raised about the sensitivity and specificity of some of these tools [52, 55].

ML-aided CAD has the potential to cross this gap with many studies reporting high sensitivity and specificity with single task questions such as pulmonary embolus detection although availability of large databases to train the algorithms is a persistent issue and hopefully projects such as the Lung Image Data consortium (LIDC). The expectation is that the Image Database Resource Initiative (IDRI), a database of annotated lung imaging, also will continue to evolve, further helping to address this issue [52].

Haematology, oncology and pathology

Haematology, oncology and histopathology are similar to radiology in that they are data heavy specialities with current practice usually involving large data sets with input from clinical examinations, electronic medical records, radiological images, histopathological images and genetic information amongst others. Coupled with this is the growing trend towards enrolling most patients in databases to track their response to treatment.

As such, these specialities have complex multidimensional data sets with many layers of interconnectivity which due to the both the large size and disparate forms of data (genetic, proteomic, clinical, biochemical, temporal, demographic, therapeutics, continuous, nominal, categorical, binary) means that using a tool that is more flexible and scalable allows for interrogation of the data that would otherwise be onerous [18, 56]. The benefits seen in radiological image analysis have been translated to histological images and a number of studies applying this ability in interesting ways.

Yu *et al.* [57] took images of lung adenocarcinoma and used image analysis software to expand the data set to provide more variables for an ML software program. The program was able to develop classifiers to automatically distinguish tumour cells and predict long-term survival, and the authors predicted that such techniques could be expanded to other cell types and contribute to precision oncology. They did this by taking 2186 histopathology specimens of benign, squamous cell and adenocarcinoma specimens and built a fully automated image segmentation program to distinguish nuclei and cytoplasm and added to this by profiling 9879 quantitative features from each image. Following this, they put the extracted data along with known patient outcomes into a variety of ML algorithms (support vector machine, naïve Bayes and random forest) to accurately predict survival outcomes superior to that by the conventional pathologic assessment based on tumour stage and grade. Similarly, but without using imaging data, Shipp *et al.* used genetic profiling linked with outcome data for diffuse large B-cell lymphoma and found a large (70 vs. 12%) difference in subgroups 5-year survival [58], potentially drastically changing treatment options for one of the most common lymphoid malignancies. This trend of AI models outperforming physician estimates is continued in that Gupta *et al.* [59] took a pre-existing cancer registry and electronic administrative records (not medical records) and found that it provided better outcome prediction than previously used models or a clinician panel.

Within Oncology, a trend towards using artificial neural networks has developed and multiple studies are demonstrating superior predictive accuracy using ML techniques in comparison with expert-based or statistical systems [34]. With bigger, linked data sets, the potential predictive ability and individualization of treatment decisions become stronger and going forward, it is imperative that these databases are maintained and expanded. As an example of this utility, Estava *et al* [60]. described the construction of a convolutional neural network (CNN) using 129 450 clinical images from a database to be able to create a model that was as sensitive and specific as board-certified dermatologists in differentiating benign versus malignant skin lesions. With the ubiquity of smartphones and relative dearth of dermatologists, the potential utility of such an application is apparent, especially in smaller, less resourced or Third World countries.

ML algorithms of relevance to medicine

There are a vast number of ML algorithms in use, the details of which are beyond the scope of this review. However, a few common ML algorithms are discussed below because of their relevance to medical research and their frequent use within the literature.

Support vector machine (SVM)

Support vector machine is a supervised learning algorithm intended to split data into two or more categories. The term 'support vector' simply refers to the margin that the algorithm uses to support its determination of whether data fall into a category or not (Fig. 3). Frequently in SVM, researchers refer to the use of 'kernels', these are mathematical tools that modify the data in certain ways to make the data more amenable to separation into categories [61]. Whilst this is straightforward for two and three-dimensional data sets, the strength of SVM is that can be used for complex data sets with many variables or dimensions. Due to its versatility, SVM has been applied to a variety of data types, from categorizing breast mammograms as having microcalcifications or not to classifying tissue and cell types based on genetic microarray expression data [62, 63].

Neural networks (NN)

Neural networks, also known as artificial NN, attempt to use multiple layers of calculations to imitate the concept of how the human brain interprets and draws conclusions from information. NN are essentially mathematical models which are designed to deal with complex and disparate information, and the nomenclature of this algorithm comes from its use of 'nodes' akin to synapses in the brain (Fig. 4) [56]. The learning process of a NN can either be supervised or unsupervised. A neural net is said to learn in a supervised manner if the desired output is already targeted and introduced to the network by training data whereas unsupervised NN have no such preidentified target outputs and the goal is to group similar units close together in certain areas of the value range.

The supervised form takes data (e.g. symptoms, risk factors, laboratory and imaging results) for training on known outcomes and searches for different combinations to find the most predictive combination of variables. NN assigns more or less weight to certain combinations of nodes to optimize the predictive performance of the trained model [64]. For example, the algorithm may find that combining blood glucose level, waist circumference and lipid levels into a node provides greater predictive accuracy of patients that will progress to become diabetic and can increase the influence or 'weight' of this node when combining it with other parts of the algorithm that look at age, body mass index, level of exercise per week and so forth to see whether this further increases predictive accuracy. By continually adjusting these relative weights of nodes, the network can progressively determine the best combination of nodes and weights that produce the optimal predictive accuracy. The lines shown in graphical descriptions of NN relate to the interdependencies of these nodes and how they influence each other. NN have been used in medical diagnosis with both conventional data sets and variables as well as data sets that use complex variables such as raw and transformed image data [65, 66], audio recordings of heart sounds and magnetic resonance spectroscopy, with significant results [67, 68].

Deep learning

Deep learning is a branch of NN characterized by multiple hidden node layers that learn representations of data by abstracting it in many ways. Where deep learning is differentiated from a simple neural network is that the number of layers of nodes is increased and the overall size of the network is larger, allowing for complex interrelationships to be represented more accurately [69]. To illustrate this, CNN, a type of feedforward NN that is designed to mimic neural processes in the brain, is frequently applied to image processing tasks [70, 71]. In these cases, it subsamples images into 'feature maps' of different aspects of the image by transforming or 'convoluting' the image into higher order features, such as contrast, lines, colour, shapes then combining or 'pooling' these features in various ways to form a depth of analysis of an image beyond simple individual pixel values or directly observed features [72]. Once a level of confidence is reached by the network that is analysing an object (such as a cross-sectional image of the brain), it can backpropagate the information to the lower level nodes. These lower level nodes look at lines, shapes and contrast to alter the algorithm to maximize the features that may represent different cross-sectional areas of the brain to determine which level of the brain is

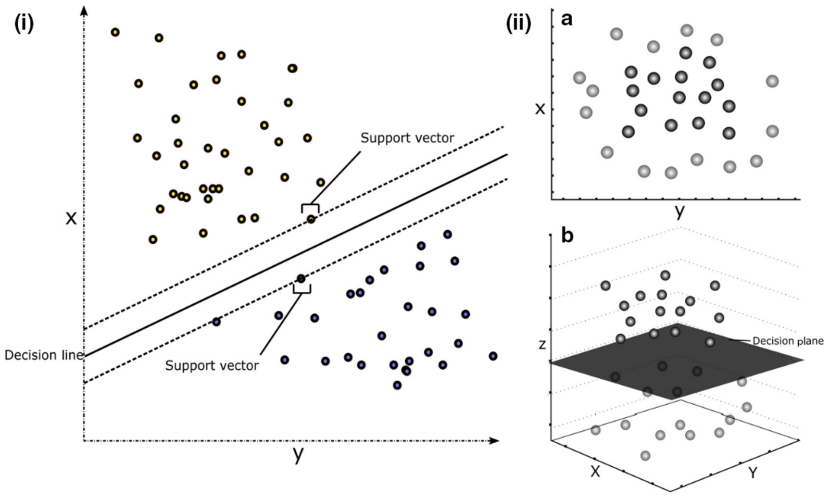

Fig. 3 *(i) Example of simple SVM graphical output. Linear separation of two-dimensional data can sometimes be impossible (iia); however, kernel transformation allows separation of data in higher dimensions (iib).*

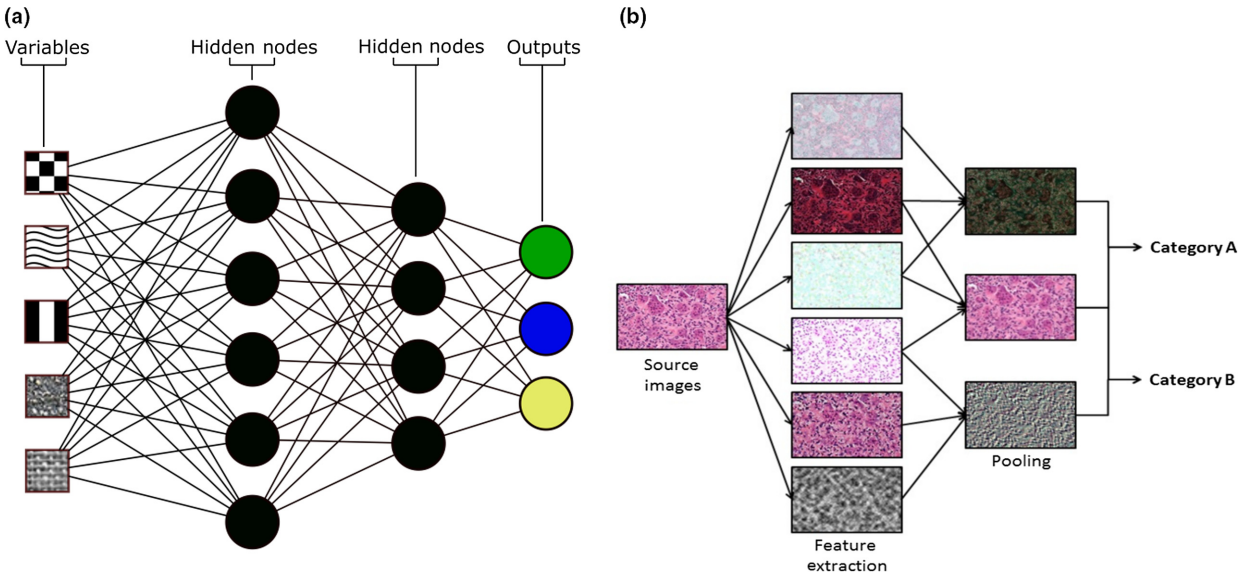

Fig. 4 *(a) Graphical representation of an archetypal ANN (b) A simplified example of image transformation for use in a CNN. Source images are manipulated to accentuate features and different combinations of these are pooled to contribute towards category definition, allowing for a novel image to be placed in the algorithms and correctly sorted to a category.*

presented in the image or if pathological characteristics are present. The key concept of providing the algorithm with a training data set to form these final conclusions is broadly similar to other models.

A current medical example of deep learning is Gulshan *et al.*'s development of a deep learning algorithm to detect diabetic retinopathy from retinal fundus photographs. They took 128 175 retinal images that had already been analysed by ophthalmologists as their training set, developed a deep learning algorithm that could analyse new images presented to it and was able to identify diabetic retinopathy with a sensitivity of 97.5% and specificity of 93.4% [73]. If results like this continue and are repeatedly validated in real-world

settings, it has huge implications for screening and service delivery.

Decision tree learning

Most clinicians will be familiar with decision trees in their practice, manifested by branching decisions based on clinical, serological or radiological data [74]. Commonly, these trees are constructed by reviewing the evidence and making suggestions based on various data sources and expert opinion. However, with the increasing data load and complexity of analysis, machine-assisted decision trees can be constructed by training the algorithm on large databases of patient cases where the outcome is known (supervised) to construct a decision tree based on which variables it can determine the highest separability on the desired categories [75]. This allows us to partition observations by using a set of simple decisions in a hierarchical basis and allows for feature selection of discriminant variables.

Latent variable models

Latent variable models are an unsupervised ML technique for investigating data for latent or unobserved variables that are not represented in the data set although may, in fact, influence the eventual outcome [76]. If we have a data set of variables of patients who develop recurrence of a type of a tumour, there is the possibility that we do not have all the data or the right representation of data to account for the outcomes for these patients and as a result, predictive models will be inaccurate.

However, knowing that there may be a latent variable allows the investigator to account for this and by manipulation of the data, use it to improve predictive accuracy or determine the strength of these latent variables. As an example, if an investigator suspects that the heterogeneity of a body of data would be best represented in smaller subgroups, latent class analysis (LCA) and principal component analysis (PCA) can be used to regress observed variables into a set of one or more latent classes that would best explain the heterogeneity in the data [77]. As an example, Chen *et al.* used data from questionnaires on 689 children regarding the presence of wheeze that was repeated over an extended period of time and by applying LCA, identified four distinct groups of wheezing patterns; early transient, early persistent, late onset and infrequent. When the data were interrogated further, the authors found that these subtypes had distinct risk factors for developing the various wheezing patterns [78]. These four wheezing patterns are the latent classes or patient phenotype; they are not directly observed as a variable in the data and identifying them is important as it allows us to focus therapies on subgroups who may benefit from them more [79]. Kutcher *et al.* [80] applied PCA to clotting factor abnormalities after acute trauma, a problem made difficult by collinearity in many clotting factors which confounds many standard regression techniques. PCA was able to identify two distinct patterns of coagulopathy (depletion and fibrinolytic) with distinct clinical patterns to them and resultant differential effect on outcomes and thus identified a possible individual diagnostic and therapeutic approach to these patients.

Assessing the machine

Much of the research in the area whilst displaying interesting results has yet to reach its full potential in clinical practice for a variety of reasons. However, one core sentiment that is difficult to overcome is that of discomfort with supplanting clinical decision-making with opaque decision tools and the problems that could arise from computer error. Notably, the QRisk2 algorithm for predicting cardiovascular risk which is in clinical use was recently incorrectly constructed and gave erroneous risk scoring, requiring many thousands of patients to be recalled and reassessed [81]. Whilst QRisk2 is not an ML algorithm, it underscores the need for oversight and validation of ML tools, especially when applied to large populations or critical decision-making for patients as once these predictive tools translate to general circulation, very few clinicians will be able to determine if there is a problem with the software.

Many ML publications fail to communicate their process effectively, providing readers with little chance to interrogate their results for error as the algorithm is not released and the data set they used is not included with the publication. In order for results to be replicated and confidence to grow in these techniques, this practice needs to change [82]. Further to this, ML is akin to conventional statistics in that it is possible for the user to make an error in the process by poor preparation of the data or inappropriately generalizing from the training sample without robust testing [83, 84].

We advocate that all physicians become familiar with ML and in particular the measures of quality so that they are able to assess the significance of information being presented to them. There are some common concepts and pitfalls that the reader should familiarize themselves with before embarking on any data analysis and Deo's article in *Circulation* further expands on this and the importance of intelligent features or variable selection and the various pitfalls of improper data input [85, 86]. Nevertheless, the most common issues lie around overfitting, underfitting, noisy data and inappropriate validation.

Assessing performance and cross-validation

For the majority of clinicians, the exposure in clinical practice to ML will be in applying the latest research and recommendations to their clinical practice. To this end, knowledge of the underlying mechanisms is not as important as understanding the performance metrics that are commonly used in assessing the quality of an ML algorithm. Some readers may be familiar with the concepts of precision, accuracy and recall and these are still used frequently. Precision refers to the repeatability of results; the degree that repeated measurement would yield the same results and is usually related to variability. Accuracy refers to how close a measurement of a variable reflects its true value and is usually related to bias. *P*-values are typically not used in ML to represent the significance of results, rather there are a few types of metrics used depending on the type of analysis undertaken, divided into problems of classification and problems of regression (Table 1).

Sensitivity (also called recall) and specificity refer to the investigators' ability to correctly identify the true-positive rate and true-negative rate, that is, those that have a disease and correctly test positive for it and those that do not have a disease and correctly test negative for it [87]. However, generalization or generalization error is a measure of how accurately an algorithm can predict outcomes for previously unseen variables and is an important metric due to an algorithm being trained on a finite number of variables and thus just relying on sensitivity and specificity results from in-sample predictions may be sensitive to sampling error [88]. For the purposes of validation, it is common practice to consciously divide data into training and testing portions. Data in the testing portion are not used to train the model so that it is possible to simulate a real-world application in dealing with previously unseen variables.

Cross-validation is a commonly used method that is an extension of this process whereby this process of segmenting the data into separate training and testing (validation) portions is repeated multiple times with different combinations of training and test data, commonly referred to as folds. This process can be repeated many times as desired to create an averaged estimation of model performance and decrease the chance of overfitting the model to the data and thus, increase the generalizability of predictions to a wider population group [89].

The receiver operating characteristic (ROC) curve is a commonly used summary statistic used in many publications as a quick way to numerically and graphically represent the performance of an algorithm (Fig. 5) for a binary classification task. It is formed by plotting the false-positive rate on the *x*-axis and the true-positive rate on the *y*-axis, enabling readers to see how using different sensitivities would affect specificity. An ideal model would have a high sensitivity with a low false-positive rate; however, naturally, for almost every test there is a trade-off at higher levels of sensitivity in increasing the number of false positives that occur. As such, it is generally better to have the apex of the curve shifted to the left and as high as possible which, consequently increases the area under the curve. This area under the receiver operating characteristic curve (AUROC) is a quick numerical representation of the overall performance; higher is better. An AUROC of 0.5 represents a classifier that performs no better than chance and an AUROC of 1 represents perfect prediction. Similarly, precision–recall curves display the trade-off between increased precision (proportion of true-positive predictions to total predicted positive) and recall (the proportion of true-positive predictions in relation to false negatives) and are represented in a similar fashion, with a higher area under the curve being desired. Both measure model performance in slightly different domains, however, a precision–recall curve is less sensitive to distortion if there is a large imbalance in class distribution (e.g. very rare disease/diagnosis) [90]. An F1 score is similar to the precision–recall curve and can be used as a summary statistic as a quick way of comparing performance; it amalgamates precision and recall of a classifier into a single score ranging from zero

Table 1 *Overview of common performance metrics used for ML models*

Indices of performance	Description
	Classification
Classification Accuracy	The ratio of correct predictions to all predictions made. Used when equal number of observations in each class and all predictions and prediction errors are equally important
AUC	The area under receiver operating characteristic curve (AUC) is used for binary classification problems. The receiver operating characteristic (ROC) curve is a graph of true-positive rate plotted against false-positive rate and the AUC is a numerical representation of the proportion of the graph that falls under this curve. An AUC of 0.5 is identical to random classification and an AUC of 1.0 represents a model that makes all predictions perfectly
Confusion Matrix	Represents classification accuracy for two or more classes calculated from a $N \times N$ matrix where N is the number of classes being predicted. Gives a variety of metrics that are informative regarding the performance of the model including overall accuracy, positive/negative predictive value, true/false-positive rate, true/false-negative rate and false omission/discovery rate
	Regression
Mean Absolute Error (MAE) and Mean Squared Error (MSE)	An indication of the magnitude of error in prediction for any given model and, as such, a number closer to zero is preferred. In regression problems, error can be positive or negative and MAE can represent the overall direction of over or under prediction. MSE squares all values so that the overall magnitude of deviation from the mean is represented, as opposed to the direction
R^2/Coefficient of Determination	R^2 is the proportion of variability in a data set that can be explained by the model and is a broad indication of how good a model will be at predicting future outcomes. It is represented on a scale of 0 to 1, with an R^2 of 0 meaning prediction is impossible based in the input variable and an R^2 of 1 means that there is no error in prediction. Intermediate values, such as 0.50 mean that you can explain 50% of what is occurring in the real data and the remainder 30% is not explained by the model you have created. Generally models should be at least above 0.6 before becoming useful

to one, with one representing perfect precision and recall.

Overfitting and underfitting

Overfitting (Fig. 6) occurs when an algorithm learns the details of the data records and noise in the training set too perfectly and the noise or random fluctuations in the training data is learned as significant concepts by the model resulting in a high variance, low bias model [91]. The result of this problem is that the insights gained from training data do not generalize to a larger and potentially more diverse data set and the model can inadvertently misclassify outliers. Similarly, underfitting refers to making a predictive model that does not utilize the pattern in the data correctly and over generalizes to subsequent data resulting in a low variance, high bias model and resultant poor predictive performance.

Dimensionality, intuition and feature engineering

It is expected that the provision of more data to an algorithm would improve accuracy. However, there is an important issue of dimensionality that needs to be considered. When an algorithm tries to group data based on similarity, this becomes increasingly difficult if extra variables are provided to it that do not contribute to classification. This can distort the pattern with noise and limits prediction and leads to what is sometimes referred to as the 'curse of dimensionality' where the inclusion of more covariables can lead to a diminution of the power of the predictive algorithm, particularly when the number of covariables exceeds the number of data points in small data sets [92].

For example, if a data set containing a large number of noncontributory variables is presented to an algorithm, it will render the algorithm useless as it will consider too much of the information when trying to separate patients into categories and increases the confusion and perplexity of the trained model. To illustrate this further, a hypothetical analysis to identify patients who will respond to antibiotics for acute appendicitis will be degraded by nonrelevant information such as eye colour, cholesterol levels or chest radiograph findings and will instead introduce noise into the data.

Whilst ML algorithms can deal with high-dimensional data, it still needs to be relevant, just as in conventional statistics, and similarly, there needs to be a degree of data preparation prior to training an ML model [83]. Thus, it is commonly accepted that 'feature engineering', feature selection and data preparation, is one of the most important steps in ML [11, 93]. To a degree, much of this can be automated within the framework of an ML software suite with components of ridge regression and LASSO (Least Absolute Shrinkage and Selection Operator – a regression analysis for variable selection and regularization) techniques to determine which variables are noncontributory but there still needs to be some intuition as regards how best to present the data to the ML software.

Determining the appropriate ML approach is ideally performed ahead of time as this will inform the data gathering approach and the required data set size. Selection of the algorithms can become a time-consuming process, especially if the data are not formatted to a form that is readily usable by the algorithm, such as transforming categorical data to ordinal or ordinal to binary. This is not a problem unique to the ML process, and many medical researchers will be familiar with it when using conventional statistical approaches but due to the high number of covariables that ML can deal with, this can be a substantial task if not anticipated. Akin to conventional statistical approaches, ML also has the possibility for user error and concern has been raised in the past regarding modelling medical systems that have no plausible biological mechanism [56, 94].

What does the future hold for machine learning in medicine?

Although there is a degree of hyperbole regarding ML potential application and its accuracy, the work done thus far gives us a degree of insight into realistic applications and with the widespread uptake of digital imaging systems, medical records and linked laboratory results across hospitals, the data stream available to researchers is increasing in quantity and quality. As an insight to this paradigm shift that is occurring, at the Radiological Society of North America annual meeting in 2017, there were 49 ML/AI exhibitors showcasing their practical application of machine learning to health care. These were not academic showcases demonstrating theory but were commercial companies already applying ML in a meaningful way, demonstrating that ML has made a transition from the academic sphere into day-to-day reality in health care enough to attract venture capital.

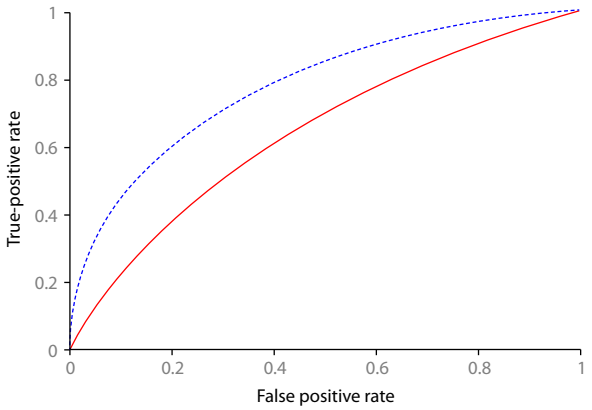

Fig. 5 *ROC curve of two hypothetical algorithms. The dashed curve has more area under its curve in comparison with the solid curve, thus can generally be thought to be better performing.*

Furthermore, there is an increasing number of publications across a variety of medical specialties trying to harness ML to improve patient care; PubMed shows in 2007 there was a total of 370 articles published on 'machine learning' and this has increased year on year by a factor of 10 to 3978 articles published in 2017. A recent controversial decision allowed Google access to the National Health Service database in the United Kingdom with its DeepMind AI to mine patient data for research purposes such as analysing retinal scan databases to detect macular degeneration and retinopathy [95, 96], radiotherapy planning for head and neck cancers and for development of an application to instantly alert healthcare providers when a patient had developed acute kidney injury.

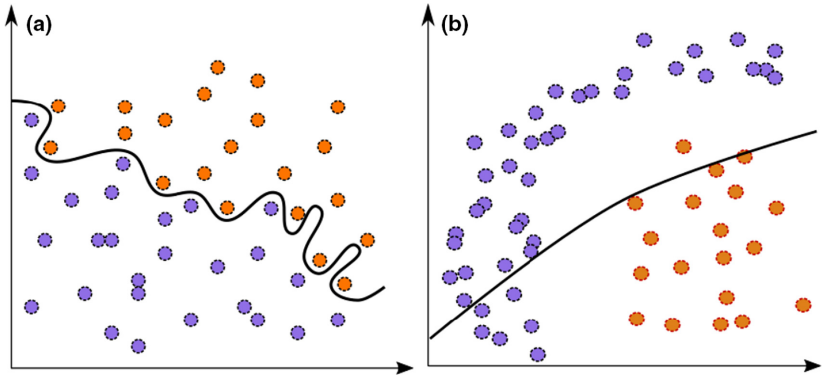

Fig. 6 *Overfitting (a) and underfitting (b) of a decision line. Overfitting results in a decision line that fits the data too exactly so that when generalized to a larger data set, miscategorization can occur. Underfitting results in a decision line that misrepresents or fails to correctly represent the pattern in the data and results in miscategorization when generalized to a larger data set.*

Google itself is a prominent developer of, and proponent of deep learning and has developed a variety of deep learning tools and teaching resources. Many of its applications have popularized the concept of deep learning and convolutional networks. It is not the only company looking to capitalize on ML in health care and there are a multitude of companies offering a variety of services in this regard. In order to facilitate this, large databases of patient data need to be used which is probably one of the reasons why medicine has fallen behind other areas in its real-world applications thus far. Data protection legislation and the fragmented nature of hospital records mean it is more difficult for researchers to utilize data for ML but as electronic care records become more commonplace and academic institutions plan for big data, these restrictions will hopefully be less of an issue in the future.

As highlighted above, there are a number of interesting and potentially practice-changing developments coming from the application of ML to healthcare data. As part of this growing trend, some clinicians such as radiologists are, in general, equal parts optimistic and concerned about how ML and AI will affect their speciality with pessimists seeing the potential for redundancy of a lot of their work if the current trajectory of success continues and thus as an extension, replacing the need for radiologists in practice; optimists see it as a useful adjunct, increasing accuracy and decreasing costly misses [97]. However, as we have shown, radiology is only part of the multitude of specialities that will be affected by ML and most specialities will see increasing use of big data and ML in their workplace as well as in print; unless clinicians understand the underpinnings, they risk being left behind. Slow, stepwise integration of ML tools is an inevitability, with automated alerts for suspected diagnosis of acute kidney injury and intracranial large vessel occlusions in stroke currently being used which will inevitably transition to more broad automated diagnosis and treatment suggestions. Being able to critically appraise a ML article that may have practice-changing implications is going to be an essential tool for a wide variety of specialities and when ML/AI begins to be introduced it behoves the clinician to be able to understand when the computer gives a treatment or diagnosis suggestion, *why and how* is it doing so; otherwise we stand to be passive users of a tool we do not understand and face redundancy in the face of an apparently superiorly performing algorithm.

Currently, the field is relatively young and many of the research groups and companies have yet to apply their products in a meaningful way in large-scale implementations. Similarly, much of the frontiers of the research in ML, whilst being shown to be theoretically powerful, has yet to make the jump to day-to-day clinical use [98]. With the growing clinical and commercial interest, this will soon change.

Conclusions

Machine learning is a continuum of the merging of computer science and statistics, and not only may

it represent the next wave in advancing modern health care, it has already arrived and is being used in real-world applications with good success in many specialties of medicine. The jump to large-scale applications and integration with general clinical practice is inevitable and the question does not seem to be if, but when.

We have reviewed the theory behind ML, explored the common ML algorithms used in medicine and discussed the potential future of ML in achieving personalized medicine. AI and ML are more likely to create computerized physician's assistant's than autonomous 'eDoctors'; however, the possibility of a 'Deep Learning Enabled' hospital is not entirely speculative and we advocate for the active development and integration of databases to facilitate this process.

Physicians should become more familiar with the basic concepts and metrics of ML, their potential applications and embrace the increasing integration of AI and ML into modern medicine.

Author Contributions

GH involved in substantial contributions to the conception and design of the work, and drafting and revising it critically for important intellectual content, and final approval of the version to be published. HKK, RC, AR and ML: involved in substantial contributions to the design of the work, and revising it critically for important intellectual content, and final approval of the version to be published. HA involved in substantial contributions to the conception and design of the work, and revising it critically for important intellectual content, and final approval of the version to be published.

Declaration of interests

The authors have no conflict of interests to declare.

References

1 Sadegh-Zadeh K. In dubio pro aegro. *Artif Intell Med* 1990; **2:** 1–3.
2 Weiss J, Kuusisto F, Boyd K, Liu J, Page D. Machine learning for treatment assignment: improving individualized risk attribution. *AMIA Annu Symp Proc* 2015; **2015:** 1306–15.
3 Weiss JC, Natarajan S, Peissig PL, McCarty CA, Page D. Machine learning for personalized medicine: predicting primary myocardial infarction from electronic health records. *AI Magazine* 2012; **33:** 33.
4 Choi E, Bahadori MT, Schuetz A, Stewart WF, Sun J. Doctor AI: predicting clinical events via recurrent neural networks. *JMLR Workshop Conf Proc* 2016; **56:** 301–18.
5 Munoz A. Machine Learning and Optimization. URL: https://www.cims.nyu.edu/~munoz/files/ml_optimization pdf [accessed 2016-03-02][WebCite Cache ID 6fiLfZvnG]. 2014.
6 Mitchell T. *Machine Learning*. New York, NY, USA: McGraw-Hill, 1997.
7 Turing AM. Computing machinery and intelligence. *Mind* 1950; **59:** 433–60.
8 Breiman L. Statistical modeling: the two cultures (with comments and a rejoinder by the author). *Statist Sci* 2001; **16:** 199–231.
9 Patel VL, Shortliffe EH, Stefanelli M *et al.* The coming of age of artificial intelligence in medicine. *Artif Intell Med* 2009; **46:** 5–17.
10 Peek N, Combi C, Marin R, Bellazzi R. Thirty years of artificial intelligence in medicine (AIME) conferences: A review of research themes. *Artif Intell Med* 2015; **65:** 61–73.
11 Guyon I, Elisseeff A. An introduction to variable and feature selection. *J Mach Learn Res* 2003; **3:** 1157–82.
12 Ringner M. What is principal component analysis? *Nat Biotech* 2008; **26:** 303–4.
13 Zhong C, Han J, Borowsky A, Parvin B, Wang Y, Chang H. When machine vision meets histology: A comparative evaluation of model architecture for classification of histology sections. *Med Image Anal* 2017; **35:** 530–43.
14 Naqa IE, Li R, Murphy MJ. *Machine Learning in Radiation Oncology: Theory and Applications*. Cham: Springer, 2015.
15 Jennifer L, Marmosh DDM. Using google reverse image search to decipher biological images. *Curr Protoc Mol Biol* 2015; **111:** 19.13.1-4.
16 Ottenbacher KJ, Linn RT, Smith PM, Illig SB, Mancuso M, Granger CV. Comparison of logistic regression and neural network analysis applied to predicting living setting after hip fracture. *Ann Epidemiol* 2004; **14:** 551–9.
17 Goldstein BA, Navar AM, Carter RE. Moving beyond regression techniques in cardiovascular risk prediction: applying machine learning to address analytic challenges. *Eur Heart J* 2017; **38:** 1805–14.
18 Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI. Machine learning applications in cancer prognosis and prediction. *Comput Struct Biotechnol J* 2015; **13:** 8–17.
19 Liang Y, Chai H, Liu X-Y, Xu Z-B, Zhang H, Leung K-S. Cancer survival analysis using semi-supervised learning method based on Cox and AFT models with L1/2 regularization. *BMC Med Genomics* 2016; **9:** 11.
20 Nordhausen K. The Elements of statistical learning: data mining, inference, and prediction, second edition by Trevor Hastie, Robert Tibshirani, Jerome Friedman. *Int Stat Rev* 2009; **77:** 482.
21 Vranas KC, Jopling JK, Sweeney TE *et al.* Identifying distinct subgroups of ICU patients: a machine learning approach. *Crit Care Med* 2017; **45:** 1607–15.
22 Wang S, Summers RM. Machine learning and radiology. *Med Image Anal* 2012; **16:** 933–51.
23 Sajda P. Machine learning for detection and diagnosis of disease. *Annu Rev Biomed Eng* 2006; **8:** 537–65.

24 Chandola V, Banerjee A, Kumar V. Anomaly detection: A survey. *ACM Comput Surv* 2009; **41:** 1–58.
25 Hauskrecht M, Valko M, Kveton B, Visweswaran S, Cooper GF. Evidence-based anomaly detection in clinical domains. *AMIA Annu Symp Proc* 2007; **2007:** 319–23.
26 Roberts S, Tarassenko L. A probabilistic resource allocating network for novelty detection. *Neural Comput* 1994; **6:** 270–84.
27 Laurikkala J, Juhola M, Kentala E, Lavrac N, Miksch S, Kavsek B, editors. Informal identification of outliers in medical data. Fifth International Workshop on Intelligent Data Analysis in Medicine and Pharmacology; 2000.
28 Chapelle O, Schlkopf B, Zien A. *Semi-Supervised Learning*. Cambridge, Massachusetts: The MIT Press, 2010; 528.
29 Timsina P, Liu J, El-Gayar OM, Shang Y, editor. Using Semi-Supervised Learning for the Creation of Medical Systematic Review: An Exploratory Analysis. 49th Hawaii International Conference on System Sciences; 2016.
30 Batmanghelich KN, Ye DH, Pohl KM, Taskar B, Davatzikos C, Adni, editors. Disease classification and prediction via semi-supervised dimensionality reduction. 2011 IEEE International Symposium on Biomedical Imaging: From Nano to Macro; 2011 March 30 2011-April 2 2011.
31 Azmi R, Norozi N, Anbiaee R, Salehi L, Amirzadi A. IMPST: a new interactive self-training approach to segmentation suspicious lesions in breast MRI. *J Med Signals Sensors* 2011; **1:** 138–48.
32 Holzinger A. Trends in interactive knowledge discovery for personalized medicine: cognitive science meets machine learning. *IEEE Intell Inform Bull* 2014; **15:** 6–14.
33 Dilsizian SE, Siegel EL. Artificial intelligence in medicine and cardiac imaging: harnessing big data and advanced computing to provide personalized medical diagnosis and treatment. *Curr Cardiol Rep* 2013; **16:** 441.
34 Cruz JA, Wishart DS. Applications of machine learning in cancer prediction and prognosis. *Cancer Inform* 2006; **2:** 59–77.
35 Menden MP, Iorio F, Garnett M *et al.* Machine learning prediction of cancer cell sensitivity to drugs based on genomic and chemical properties. *PLoS ONE* 2013; **8:** e61318.
36 Guan WJ, Jiang M, Gao YH *et al.* Unsupervised learning technique identifies bronchiectasis phenotypes with distinct clinical characteristics. *Int J Tuberc Lung Dis* 2016; **20:** 402–10.
37 Howard R, Rattray M, Prosperi M, Custovic A. Distinguishing asthma phenotypes using machine learning approaches. *Curr Allergy Asthma Rep* 2015; **15:** 38.
38 Salgado CM, Vieira SM, Mendonça LF, Finkelstein S, Sousa JMC. Ensemble fuzzy models in personalized medicine: Application to vasopressors administration. *Eng Appl Artif Intell* 2016; **49:** 141–8.
39 Weng SF, Reps J, Kai J, Garibaldi JM, Qureshi N. Can machine-learning improve cardiovascular risk prediction using routine clinical data? *PLoS ONE* 2017; **12:** e0174944.
40 Kirchmair J, Goller AH, Lang D *et al.* Predicting drug metabolism: experiment and/or computation? *Nat Rev Drug Discov* 2015; **14:** 387–404.
41 Gawehn E, Hiss JA, Schneider G. Deep learning in drug discovery. *Mol Inform* 2016; **35:** 3–14.
42 Stanfield Z, Coskun M, Koyuturk M. Drug response prediction as a link prediction problem. *Sci Rep* 2017; **7:** 40321.
43 Kassahun Y, Yu B, Tibebu AT *et al.* Surgical robotics beyond enhanced dexterity instrumentation: a survey of machine learning techniques and their role in intelligent and autonomous surgical actions. *Int J Comput Assist Radiol Surg* 2016; **11:** 553–68.
44 Garg AX, Adhikari NJ, McDonald H *et al.* Effects of computerized clinical decision support systems on practitioner performance and patient outcomes: A systematic review. *JAMA* 2005; **293:** 1223–38.
45 Kiranantawat K, Sitpahul N, Taeprasartsit P *et al.* The first Smartphone application for microsurgery monitoring: SilpaRamanitor. *Plast Reconstr Surg* 2014; **134:** 130–9.
46 Automation and anxiety. The Economist. 2016 June 25th 2016.
47 Garcia-Lorenzo D, Lecoeur J, Arnold DL, Collins DL, Barillot C. Multiple sclerosis lesion segmentation using an automatic multimodal graph cuts. *Med Image Comput Comput-Assist Interv* 2009; **12:** 584–91.
48 Cocosco CA, Kedenburg G, Niessen WJ, Thoms H. A method a system and a computer program for segmenting a structure associated with a reference structure in an image. Google Patents; 2007.
49 Kedenburg G, Cocosco CA, Köthe U, Niessen WJ, Vonken E-jPA, Viergever MA, editors. Automatic cardiac MRI myocardium segmentation using graphcut 2006.
50 Cheng HD, Cai X, Chen X, Hu L, Lou X. Computer-aided detection and classification of microcalcifications in mammograms: a survey. *Pattern Recogn* 2003; **36:** 2967–91.
51 Summers RM, Yao J, Pickhardt PJ *et al.* Computed tomographic virtual colonoscopy computer-aided polyp detection in a screening population. *Gastroenterology* 2005; **129:** 1832–44.
52 Chan HP, Hadjiiski L, Zhou C, Sahiner B. Computer-aided diagnosis of lung cancer and pulmonary embolism in computed tomography-a review. *Acad Radiol* 2008; **15:** 535–55.
53 Rahman MM, Bhattacharya P, Desai BC. A framework for medical image retrieval using machine learning and statistical similarity matching techniques with relevance feedback. *IEEE Trans Inf Technol Biomed* 2007; **11:** 58–69.
54 Akgul CB, Rubin DL, Napel S, Beaulieu CF, Greenspan H, Acar B. Content-based image retrieval in radiology: current status and future directions. *J Digit Imaging* 2011; **24:** 208–22.
55 Taylor P, Potts HWW. Computer aids and human second reading as interventions in screening mammography: Two systematic reviews to compare effects on cancer detection and recall rate. *Eur J Cancer* 2008; **44:** 798–807.
56 Schwarzer G, Vach W, Schumacher M. On the misuses of artificial neural networks for prognostic and diagnostic classification in oncology. *Stat Med* 2000; **19:** 541–61.
57 Yu K-H, Zhang C, Berry GJ *et al.* Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features. *Nat Commun* 2016; **7:** 12474.
58 Shipp MA, Ross KN, Tamayo P *et al.* Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nat Med* 2002; **8:** 68–74.
59 Gupta S, Tran T, Luo W *et al.* Machine-learning prediction of cancer survival: a retrospective study using electronic administrative records and a cancer registry. *BMJ Open* 2014; **4:** e004007.

60 Esteva A, Kuprel B, Novoa RA *et al.* Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 2017; **542:** 115–8.
61 Cortes C, Vapnik V, editors. Support-vector networks. *Mach Learn* 1995; **20:** 273–97.
62 El-Naqa I, Yang Y, Wernick MN, Galatsanos NP, Nishikawa RM. A support vector machine approach for detection of microcalcifications. *IEEE Trans Med Imaging* 2002; **21:** 1552–63.
63 Furey TS, Cristianini N, Duffy N, Bednarski DW, Schummer M, Haussler D. Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics* 2000; **16:** 906–14.
64 Abdi H. A neural network primer. *J Biol Syst* 1994; **02:** 247–81.
65 Amato F, López A, Peña-Méndez EM, Vaňhara P, Hampl A, Havel J. Artificial neural networks in medical diagnosis. *J Appl Biomed* 2013; **11:** 47–58.
66 Tate AR, Underwood J, Acosta DM *et al.* Development of a decision support system for diagnosis and grading of brain tumours using in vivo magnetic resonance single voxel spectra. *NMR Biomed* 2006; **19:** 411–34.
67 Uğuz H. A biomedical system based on artificial neural network and principal component analysis for diagnosis of the heart valve diseases. *J Med Syst* 2012; **36:** 61–72.
68 Brougham DF, Ivanova G, Gottschalk M, Collins DM, Eustace AJ, O'Connor R, Havel J. Artificial neural networks for classification in metabolomic studies of whole cells using 1H nuclear magnetic resonance. *J Biomed Biotechnol* 2011; **2011:** 8.
69 Lee JG, Jun S, Cho YW *et al.* Deep learning in medical imaging: general overview. *Korean J Radiol* 2017; **18:** 570–84.
70 Egmont-Petersen M, de Ridder D, Handels H. Image processing with neural networks—a review. *Pattern Recogn* 2002; **35:** 2279–301.
71 Aizenberg I, Aizenberg N, Hiltner J, Moraga C, Meyer zu Bexten E. Cellular neural networks and computational intelligence in medical image processing. *Image Vis Comput* 2001; **19:** 177–83.
72 LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature* 2015; **521:** 436–44.
73 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:** 2402–10.
74 Thomas JB. Advanced cardiac life support (ACLS) algorithms. A powerful decision tree for management of cardiac arrest victims. *Can Crit Care Nurs J* 1989; **6:** 12–9.
75 Podgorelec V, Kokol P, Stiglic B, Rozman I. Decision trees: an overview and their use in medicine. *J Med Syst* 2002; **26:** 445–63.
76 Cai L. Latent variable modeling. *Shanghai Arch Psychiat* 2012; **24:** 118–20.
77 Miettunen J, Nordström T, Kaakinen M, Ahmed AO. Latent variable mixture modeling in psychiatric research – a review and application. *Psychol Med* 2015; **46:** 457–67.
78 Chen Q, Just AC, Miller RL *et al.* Using latent class growth analysis to identify childhood wheeze phenotypes in an urban birth cohort. *Ann Allergy Asthma Immunol* 2012; **108:** 311–5.
79 Wraith D, Wolfe R. Classifying patients by their characteristics and clinical presentations; the use of latent class analysis. *Respirology* 2014; **19:** 1138–48.
80 Kutcher ME, Ferguson AR, Cohen MJ. A principal component analysis of coagulation after trauma. *J Trauma Acute Care Surgery* 2013; **74:** 1223–30.
81 Agency MaHpR. MHRA information on TPP and QRISK®2 2016 [Available from: https://www.gov.uk/government/news/mhra-information-on-tpp-and-qrisk2.
82 Dreiseitl S, Ohno-Machado L. Logistic regression and artificial neural network classification models: a methodology review. *J Biomed Inform* 2002; **35:** 352–9.
83 Domingos P. A few useful things to know about machine learning. *Commun ACM* 2012; **55:** 78–87.
84 Nuzzo R. Scientific method: statistical errors. *Nature* 2014; **506:** 150–2.
85 Lemm S, Blankertz B, Dickhaus T, Müller K-R. Introduction to machine learning for brain imaging. *NeuroImage* 2011; **56:** 387–99.
86 Deo RC. Machine learning in medicine. *Circulation* 2015; **132:** 1920–30.
87 Altman DG, Bland JM. Statistics notes: diagnostic tests 1: sensitivity and specificity. *BMJ* 1994; **308:** 1552.
88 Wolpert DH. Stacked generalization. *Neural Networks* 1992; **5:** 241–59.
89 Arlot S, Celisse A. A survey of cross-validation procedures for model selection. *Statist Surv* 2010; **4:** 40–79.
90 Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS ONE* 2015; **10:** e0118432.
91 Obermeyer Z, Emanuel EJ. Predicting the future — big data, machine learning, and clinical medicine. *N Engl J Med* 2016; **375:** 1216–9.
92 Friedman JH. On Bias, Variance, 0/1—Loss, and the curse-of-dimensionality. *Data Min Knowl Disc* 1997; **1:** 55–77.
93 Bermingham ML, Pong-Wong R, Spiliopoulou A *et al.* Application of high-dimensional feature selection: evaluation for genomic prediction in man. *Sci Rep* 2015; **5:** 10312.
94 Strasak AM, Zaman Q, Pfeiffer KP, Gobel G, Ulmer H. Statistical errors in medical research–a review of common pitfalls. *Swiss Med Wkly* 2007; **137:** 44–9.
95 Hawkes N. NHS data sharing deal with Google prompts concern. *BMJ* 2016; **353:** i2573.
96 Shah H. The DeepMind debacle demands dialogue on data. *Nature* 2017; **547:** 259.
97 Recht M, Bryan RN. Artificial intelligence: threat or boon to radiologists? *J Am College Radiol* 2017; **14:** 1476–80.
98 Clifton DA, Niehaus KE, Charlton P, Colopy GW. Health informatics via machine learning for the clinical management of patients. *Yearbook Med Inform* 2015; **10:** 38–43.

Correspondence: Guy S. Handelman, Department of Radiology, Royal Victoria Hospital, Grosvenor Road, Belfast, UK.
(fax: +4402890635849; e-mail: guyhandelman@rcsi.ie)

