Biophysical Reviews (2019) 11:111–118
https://doi.org/10.1007/s12551-018-0449-9

REVIEW

Machine learning: applications of artificial intelligence to imaging and diagnosis

James A. Nichols[1] · Hsien W. Herbert Chan[2,3] · Matthew A. B. Baker[4]

Received: 8 May 2018 / Accepted: 14 August 2018 / Published online: 4 September 2018

Abstract

Machine learning (ML) is a form of artificial intelligence which is placed to transform the twenty-first century. Rapid, recent progress in its underlying architecture and algorithms and growth in the size of datasets have led to increasing computer competence across a range of fields. These include driving a vehicle, language translation, chatbots and beyond human performance at complex board games such as Go. Here, we review the fundamentals and algorithms behind machine learning and highlight specific approaches to learning and optimisation. We then summarise the applications of ML to medicine. In particular, we showcase recent diagnostic performances, and caveats, in the fields of dermatology, radiology, pathology and general microscopy.

Keywords Machine learning · Microscopy · Imaging · Artificial intelligence · Computer vision · Dermatology · Radiology

Fundamentals of machine learning

Machine learning (ML) is an umbrella term that refers to a broad range of algorithms that perform intelligent predictions based on a data set. These data sets are often large, perhaps consisting of millions of unique data points. Recent progress in machine learning has attained what appears to be a human level of semantic understanding and information extraction, and sometimes the ability to detect abstract patterns with greater accuracy than human experts.

Extending upon classical techniques in statistical modelling, modern machine learning has emerged as a powerful tool due to the vastly increased volumes of data, exponential growth in computational power and advances in algorithm design, driven by the needs of web industries.

This article is part of a Special Issue on 'Big Data' edited by Joshua WK Ho and Eleni Giannoulatou

✉ Matthew A. B. Baker
matthew.baker@unsw.edu.au

[1] Laboratoire Jacques-Louis Lions, Sorbonne Université, Paris, France

[2] Centenary Institute, The University of Sydney, Sydney, Australia

[3] Department of Dermatology, Royal Prince Alfred Hospital, Camperdown, Australia

[4] School of Biotechnology and Biomolecular Sciences, University of New South Wales, Kensington, Australia

A wide variety of machine learning algorithms, which we typically refer as a *model*, are in use today. The choice of a particular model for a given problem is determined by the characteristics of the data as well as the type of desired outcome. A primary consideration is the number of unique data points. Large data sets, of the order of 10^6 unique data points, mean more exotic deep learning algorithms may be suitable. Fewer data points indicate that robust classical techniques like linear regression, or decision-tree methods which segment data sets into regions according to fixed rules, are likely to perform better. Care must be taken to tailor the approach to the characteristics of the data, whether it is a collection of images, a time-series signal or general descriptive data.

Another choice is between *supervised learning* and *unsupervised learning*. Supervised learning, as the name suggests, involves teaching the model with a collection of input data that has the correct output already associated with it. Supervised learning is more broadly used in image classification tasks. Unsupervised learning is where a model trains itself on data, in a sense. Typically, this may involve tasks like cluster detection or various forms of pattern recognition. Google's AlphaGo Zero (Silver et al. 2017) is an advanced example of unsupervised learning, where adversarial neural network models competed to learn winning moves in the game of Go. After only 3 days of reinforcement learning, AlphaGo Zero surpassed the level of the first supervised learning model from 2016, AlphaGo Lee; and after 40 days of self-

learning, it became the best Go player of all time, man or machine, and all with no human intervention.

The standard choice in model outcomes is between *classification* and *regression.* Classification is a prediction (Fig. 1), from data, of a qualitative label (e.g. labelling whether an image shows a cat or a dog), and regression is the prediction of a continuous variable (e.g. given the height of an individual, how much are they likely to weigh?).

We introduce the following notation: Given N items of data x_i and associated outcomes y_i, where i is the index from 1 to N; we choose a model, which is essentially a function $f(x, \theta)$ where x is the input data and θ represents the model parameters (θ represents a collection of parameters, there may be many more than one). The goal is to iterate towards the parameters θ that give us predicted outcomes $\hat{y}_i = f(x_i, \theta)$ that are as close as possible to y_i. The model can now be used to make predictions $y = f(x, \theta)$ with new and previously unseen input data x.

Linear regression

Linear regression is the simplest form of machine learning. We assume a linear function $f(x, \theta) = \beta x + m$ as our model, where the parameter set $\theta = (\beta, m)$ contains the slope β and intercept m of the line. In ML parlance, the calculation of slope and intercept is the *training* of the model.

The slope β and intercept m are typically found with a simple closed-form calculation, for example calculating linear least-squares yields a result that is known to minimise the sum of squares of the difference $(y_i - \hat{y}_i)^2$. The simplicity of this training procedure sits in contrast with modern deep neural networks where the model parameters can number in the millions, and an iterative method is used to search the multi-dimensional parameter space until the predictions $\hat{y}_i$ on the training data are deemed to fit y_i closely enough.

Supervised learning and classifiers

Most models of relevance to medical imaging are classifier algorithms that are trained in a supervised manner. Supervised learning, as the name suggests, involves a teacher. This role is played by a data set where each data point x_i (which could be an image or a signal) has the associated outcome y_i, specifying which of K possible classifications x_i belong to (e.g. that the image x_i represents a cat, dog or goat, in which case $K = 3$).

One splits the data in two portions; the first being the *training data*, used to find the parameters $\hat{\theta}$ that produce model results $\hat{y}_i$ closest to y_i. The other portion, the *test data*, is used primarily to assess the performance of our model but should not be used to influence our model parameters. Some practitioners also reserve a portion of *validation data*, used to select the best performer of a collection of models that may use completely different algorithms or training methods.

The typical output of a classifier model is a vector $\hat{y}_i = (\hat{y}_{i,1}, \ldots, \hat{y}_{i,K})$ that represents the probability that x_i is of class 1 to K, and the values sum to 1. The training data y_i will be vector of 0 except for the correct label K for which $y_{i,\,K} = 1$. We note that in many examples of clinical machine learning, we want a simple model that can perform *binary classification* where $K = 1$, e.g. whether a skin image is cancerous or not. The output of the model with given a single probability $\hat{y}_i$ that is between 0 and 1, we take a value about 0.5 to indicate "positive" and less that 0.5 to indicate "negative".

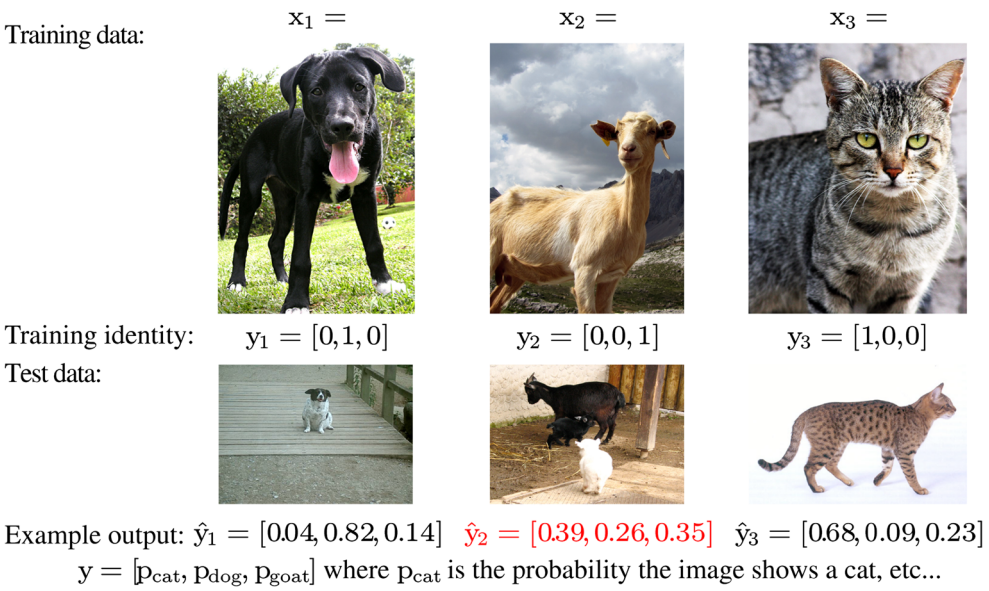

Fig. 1 Classification. An example of the individual inputs and probabilistic outputs of a classifier model. The system comprises of a ternary classifier where an image can be either a cat, a dog or a goat. In this example, the system is trained on a set of images which are labelled data as cat ($y_3 = [1, 0, 0]$), dog ($y_1 = [0, 1, 0]$) or goat ($y_2 = [0, 0, 1]$). The classifier then runs on a new test set of data,where it correctly identifies the dog and the cat but erroneously classifies the goat image as a dog

Publicly available data sets such as ImageNet, launched in 2009 (now with ~14 million images) (Jia Deng et al. 2009; Russakovsky et al. 2015) and Google's Open Images Dataset (released in 2016, ~9 million images) allow researchers to train and test new variants of deep learning models. Both are human-annotated with labels of what the image contains, and the latter Open Image set allows for multiple labels per image with location information (bounding boxes) included. The introduction of ImageNet facilitated a sharp acceleration in machine learning for image classification, including AlexNet (Krizhevsky et al. 2012) which won the 2012 ImageNet challenge by a wide margin and popularised the use of convolutional neural networks for image classification.

Deep learning and neural networks

Neural networks are now well established. Early concepts were proposed in the 1960s, but research activity picked up in the 1980s (Y. LeCun 1988; Parker 1985) and then again in the 2000s as data availability grew. Broadly speaking, neural nets mimic their biological counterparts, passing data through a web of nodes organised in interconnected layers, where the data is multiplied by a series of different weights between each node, until a final layer will give us the regression or classification answer we seek. The critical component in a neural network is training such that good weights are found in the mappings between nodes.

Convolution neural networks (CNNs) are a variant of neural nets where the first few layers of the neural net compare each part of an image against some small sub-image. Each node holds some small feature and its output to the next layer depends on how much a part of the image resembles that feature (performed by convolution). After the convolution layers, a standard fully connected neural net follows that performs the classification of the overall image (the *pooling layer*). The convolution features and the network weights are all trainable parameters.

The primary advance brought by neural net learning algorithms was the ability to perform classification without defined "feature detection", for example, without having to declare specific parameters to recognise a human face (e.g. two dark eyes, nose in the middle). Instead, neural nets are free to find their own rules of how to decompose features. This is known as *feature learning* or *representation learning*.

Loss functions and the training a model

The training of a model requires a *loss function* $L(\hat{y}_i, y_i)$, sometimes called an *objective*, *cost* or *fitness* function. It assesses how closely the model prediction $\hat{y}_i$ fits the correct value y_i, and a smaller value indicates a better fit. The training procedure then seeks to minimise the total loss $R(\theta) = \sum_{i=1}^{N} L(\hat{y}_i, y_i)$.

The most common loss function used in deep classification models is *cross entropy*, defined as $L_{CE}(\hat{y}_i, y_i) = -\sum_{k=1}^{K} y_{i,k} \log \hat{y}_{i,k}$. Mean squared error (MSE) is also used, defined as $L_{SS}(\hat{y}_i, y_i) = \sum_{k=1}^{K} (y_{i,k} - \hat{y}_{i,k})^2$, although it is more common for regression problems. Minimisation of the cross entropy method is optimal in situations which require accurate estimation of small probabilities, and is suited to predict class probabilities, whereas MSE is more suited to predicting values.

We now want to find a set of parameters θ that minimise $R(\theta)$. As θ is usually a large set of parameters, an exhaustive search through the space of all possible θ to find the value that gives the lowest value for $R(\theta)$ is computationally infeasible.

In place of an exhaustive search, we typically iteratively search for a suitable θ via a method known as *stochastic gradient descent* (Bottou 2010). For this, we examine the values of the $R(\theta)$ and θ in the direction in it decreases the fastest, with a random shuffling at each step to avoid getting stuck in local minima where the model may appear as well-trained, but a globally better candidate exists elsewhere (Fig. 2).

The use of *backpropagation* (LeCun 1988; Rumelhart et al. 1986), a method to efficiently calculate the gradient of $R(\theta)$ for neural networks, has contributed greatly to the success of neural networks in machine learning. However, it is still not uncommon for the training stage of a large model to take hours or days to complete.

Often in simple models, the initial parameters θ are simply taken to be random numbers. In modern convolution neural nets, the parameters are instead often taken to be "pretrained" parameters that have already roughly been trained on a standard image data set like ImageNet, a practice which is called

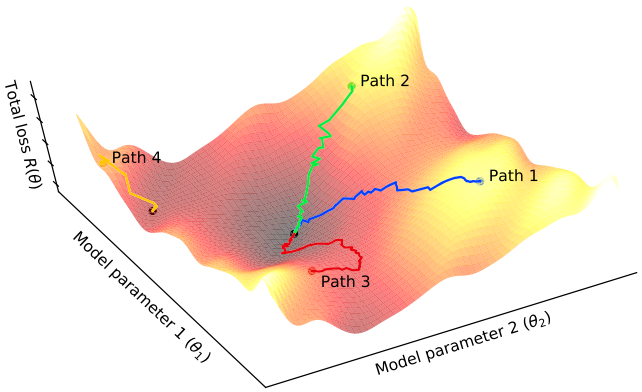

Fig. 2 Stochastic gradient descent. An illustration of stochastic gradient descent for an abstract function $R(\theta_1, \theta_2)$. Three initial estimates (path 1, blue; path 2, green; path 3, red) are shown that lead to discovery of the same global minimum (paths 1–3). However, one initial estimate and its subsequent path (path 4, yellow) lead to an erroneous local minimum

transfer learning. This enables vastly improved training times for complicated neural net algorithms.

Evaluation, overfitting and the bias-variance trade-off

When the training is finished, the effectiveness of a model is evaluated using the reserved test data. Here, typical measures such as *sensitivity* or *specificity* are typically used for binary classifiers. If a test set has P total "positive" and N total "negative" data points, then TP represents the number of true positive predictions and TN the number of true negative predictions by our classifier algorithm. Sensitivity is defined as $\frac{TP}{P}$ and specificity as $\frac{TN}{N}$. A plethora of alternative tests are available and are important to consider depending on the structure of the data, for example if positive data points are likely to be very rare (Fawcett 2006).

A common problem is *overfitting*, where the model is trained too specifically to the training data and may not subsequently perform so well in predictive use in the field (Fig. 3). This can be a particularly acute problem when a model has a much larger number of parameters than the number of available training data points. This is akin to fitting a large complex polynomial to a limited number of data points—providing a perfect, but incorrect, fit. Similarly, when given a large enough neural net, it is possible to train such that perfect results are obtained on the trained data set, but with incorrect predictive or diagnostic capacity.

There are a few strategies to avoid overfitting. One can apply or penalties for increasing model complexity, or track evaluation metrics vs time in order to stop training earlier before a model is overfit. Other common strategies include *cross-validation* (Kohavi 1995). For example in *K-fold cross-validation*, the training data is split in to K (usually 10) segments; the model is trained K times, each time with a different segment missing, and then out of the K sets of model parameters, then the one that gives the best results over all data is then chosen. Other methods to avoid overfitting include *bootstrap aggregating* or *bagging* (Breiman 1996) and the *dropout method* for neural nets (Srivastava et al. 2014).

The converse problem to overfitting is underfitting. This describes the situation where the model lacks some of the relevant assumptions or complexity to accurately reflect the physical system that the data comes from. The model will make inaccurate predictions with new data. Linear models as discussed above, for example, may be too simple for many data sets. Alternatively, under-trained neural nets may exhibit very simple output behaviour that is considered underfit and inaccurate. There are reasonably reliable metrics to detect underfitting, so it is often considered less of a problem than overfitting (Hastie et al. 2009).

An important statistical concept here is that an underfit model will display high *bias* but low *variance*, and vice-versa for an overfit model. These terms have precise mathematical definitions but cannot be calculated directly, only approximated. Loosely speaking, bias measures the error of our model vs the "true" underlying model, and variance measures how much our model will change if it were given different training data (ideally not at all!). Typically, there is a trade-off between bias and variance, and the ideal model minimises their sum, at which point the model may be considered detailed enough to capture the physical reality, but simple enough to not be excessively specific to our training data.

Clinical applications

The applications of machine learning to clinical medicine align strongly with computer vision tasks of detection, segmentation and classification; for example, the detection of the presence or absence of metastases on histological sections, segmentation of radiological images into known anatomical correlates and the classification of images into certain diagnostic categories (Fig. 4).

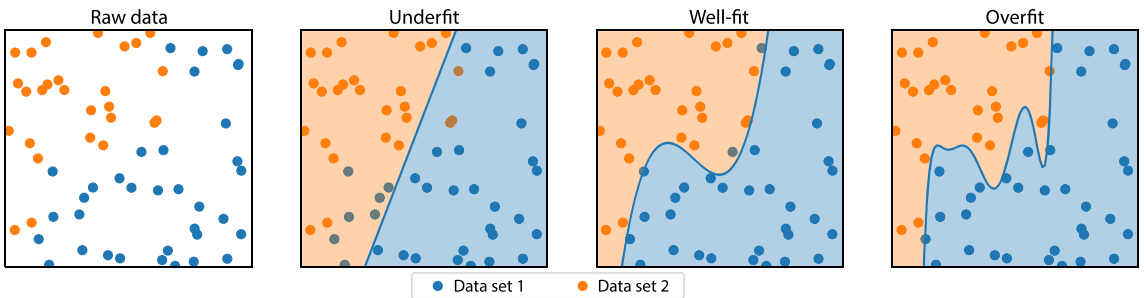

Fig. 3 Underfitting, overfitting and the bias-variance trade-off. An illustration demonstrating a classification problem (segmenting two data sets, blue and orange). **a** The raw data set. **b** An example of underfitting, where a too simple separation has resulted in misclassifying some members of each data set. **c** An example of a well-fit classifier which correctly separates the data sets and classifies correctly nearly all members of both data sets, without too complex a model. **d** An overfit classifier, which correctly identifies all the members of each data sets but is overly complex, has high variance, and incorrectly separates the space

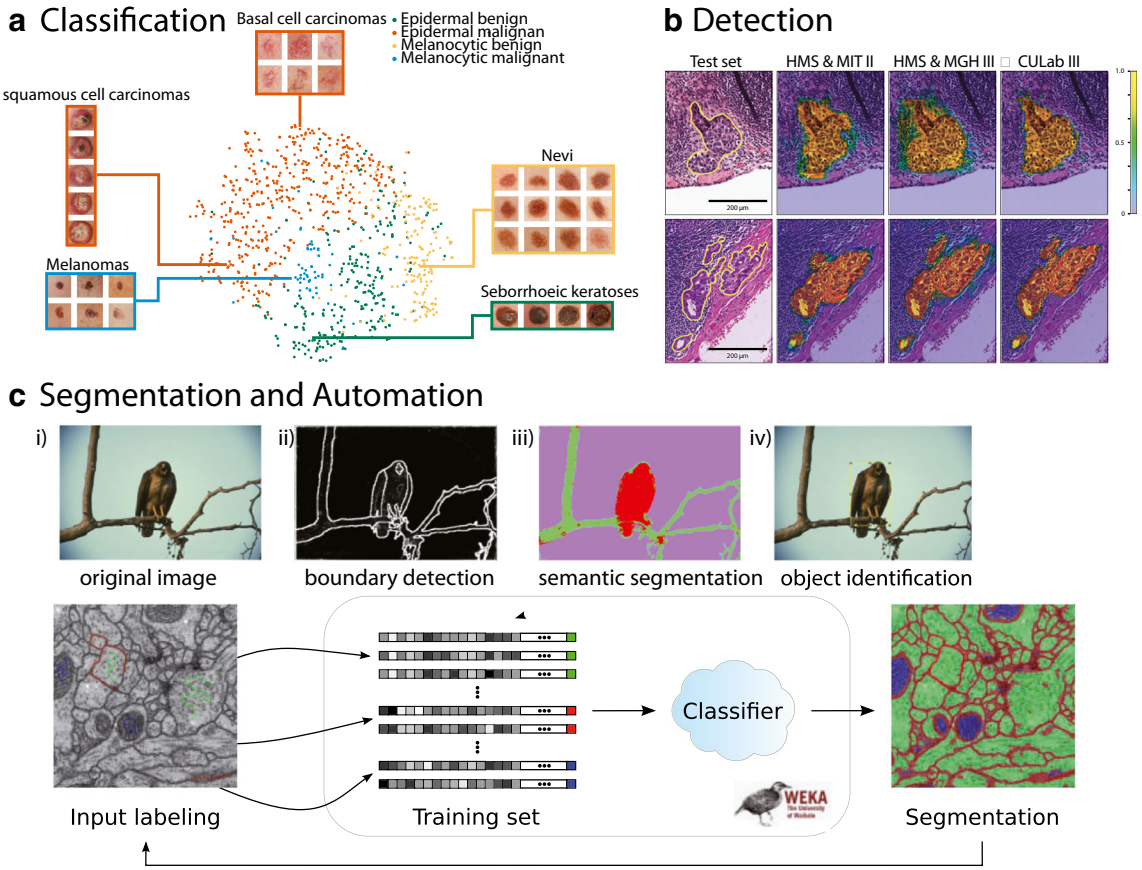

Fig. 4 Applying machine learning to medical imaging. **a** Demonstration of classification and clustering in the final hidden layer of the convolutional neural network in biopsy-proven photographic test sets. Coloured point clouds represent different disease categories, insets show images corresponding to various points. Effective clustering of similar diagnosis can be observed by eye and measured statistically with ROC curves (taken from Esteva et al. 2017). **b** Computer-aided diagnosis of lymph node metastases in women with breast cancer. The colour scale bar (top right) indicates the probability for each pixel to be part of a metastatic region. The top and bottom rows show two annotated micrometastatic regions in whole-slide images of haematoxylin and eosin-stained lymph node tissue sections taken from the test set of Cancer Metastases in Lymph Nodes Challenge 2016 (CAMELYON 16) dataset. Second through fourth columns show probability maps from each team overlaid on the original image: HMS, Harvard Medical School; MIT, Massachusetts Institute of Technology; MGH, Massachusetts General Hospital and CULab, Chinese University Lab. The top two deep learning-based systems, from the teams HMS and MITII and HMS and MGH III outperformed all the pathologists without time constraint in this study (taken from Beijnordi et al. 2017). **c** Trainable WEKA segmentation pipeline for pixel classification. Image features are first extracted using native non-machine learning methods inside the imaging software Fiji [ref] (**i**). One example of such a method is edge detection (**ii**) through the Canny edge detection which is based on analysis of gradients. Next, a WEKA (Hall et al. 2009) learning scheme is trained on a set of pixel samples represented as feature vectors (from various image features), and the user provides iterative and interactive feedback to correct or add labels. This is then used for semantic segmentation of the image (**iii**) and finally object identification (**iv**). An example pipeline showing a serial section from transmission electron microscopy of Drosophila larva ventral nerve cord with pixels divided into three classes: membrane, mitochondria and cytoplasm (bottom) (taken from Arganda-Carreras et al. 2017)

Detecting features in images, such as edge detection of various boundaries, originated in the 1970s using algorithms based on pixel intensity and gradients in intensity, with thresholds set by the user (Spontón and Cardelino 2015). However recently, the advent in particular of deep learning methods and artificial neural networks, has allowed for an end-to-end process where the features determined by the algorithm are used to minimise a loss function through an iterative process thus removing the need for user input (Chartrand et al. 2017).

The success of artificial neural networks has been enabled by the increasing availability of large databases of raw data with attached diagnoses (labelled data), improvements in computer hardware and improvements in network architecture and training techniques (Chartrand et al. 2017). Recently, the application of these deep learning algorithms has led to significant breakthroughs in the field of dermatology, radiology, ophthalmology and cardiology.

Classification

A landmark paper in the application of neural networks to clinical diagnosis, and the collaboration between computer scientists and dermatologists, was the Nature paper of

Esteva et al. (2017). They tested two binary classifications of keratinocytic carcinomas vs benign seborrheic keratoses and malignant melanomas vs benign nevi, based on viewing biopsy-proven clinical images. Their deep CNN outperformed 21 board-certified dermatologists. While there have been previous attempts at applying CNN to computer-aided diagnosis, this application of CNN to dermatology by Esteva et al. differed in the architecture of their algorithms, the number of images used to train the algorithm, the use of a taxonomy-based partitioning algorithm and no requirement for preprocessing images. Additionally, other than resizing test images to 299 × 299 pixels, the images were non-standardised. The architecture of the CNN was based on GoogleNet Inception v3 (Szegedy et al. 2015) which was pretrained on approximately 1.28 million images consisting of 1000 object categories.

Through a process of transfer learning, the CNN was then trained on a dermatology image database consisting of 127,463 training and validation images including 1942 biopsy-labelled test images across 757 disease classes and 2032 tree-structured taxonomy of disease labels which were derived by a disease partitioning algorithm (Pan and Yang 2010). Three diagnostic tasks were assessed: keratinocyte carcinoma classification, melanoma classification and melanoma classification using dermoscopic images of skin lesions under × 10 magnification.

The key elements used in this comparison were (1) sensitivity, the true positive rate and (2) specificity, the true negative rate. Since ML methods naturally use probabilistic estimation, typically a threshold, t, is set such that the prediction $\hat{y}$ for any image is $\hat{y} = p \geq t$, where p is the probability of malignancy per image (in the Esteva et al. case). As t is varied from between 0 and 1, an algorithm can have perfect sensitivity (everything is classified as malignant, thus true positive rate is 100%, but many non-malignant images are wrongly classified) or perfect specificity (true negative rate is 100%, but many malignant images are wrongly classified as non-malignant). By plotting sensitivity vs specificity for the algorithm for all values of t, a CNNs performance is able to be measured across all t, and is reported as the area under the curve (AUC) as a fraction between 0 and 1. A clinician's diagnosis is represented as a single point (their sensitivity/specificity ratio), and if they lie below the line for the algorithm, they have been outperformed by the CNN. Esteva et al. achieved both high sensitivity and specificity (AUC > 0.91).

A recent progression to the above method in melanoma classification has been the combination of algorithms aggregated via a machine learning fusion algorithm. The 2016 International Skin Imaging Collaboration International Symposium on Biomedical Imaging challenge sets the task to diagnose melanoma based on dermoscopic images. While individual algorithms performed comparably to clinicians, a fusion algorithm (greedy fusion, ROC = 0.86) outperformed clinicians (ROC = 0.71, $p = 0.001$) (Marchetti et al. 2018).

A field highly suited to classification applications of machine learning algorithms is that of radiology where large electronic databases of standardised images with labelled diagnoses already exist, and computer-aided diagnosis (CAD) is a rapidly emerging field. Deep convolutional neural networks have recently been applied to many areas of radiography for a variety of different tasks and have been extensively reviewed in Litjens, Medical Image Analysis, 2017 (Litjens et al. 2017). A number of different features influence the design of the algorithm and its performance including the data set used for training and assessment, the type of CNN used, the number of levels within the CNN (whether the CNN is pretrained or not), the integration of transfer learning or random initialisation processes and also augmentation either through human input or fusion algorithms.

Detection

Pathology underpins and informs medical treatment across multiple fields. A correct diagnosis refers to the presence, absence and severity grade of pathology and determines decision making in prognosis, risk assessment and treatment. Pathology is a natural target for optimisation and automation via machine learning since large image-based datasets exist. ML is frequently applied to image data generated by histopathology, fundoscopy and radiography.

As part of the recent 2017 IEEE international symposium on Biomedical Imaging, a challenge was set for computer-aided diagnosis of lymph node metastases in women with breast cancer (CAMELYON 16). Large datasets comprised of whole-slide images of approximately 200,000 × 100,000 pixels were used, and the complete training data set was over 500 GB consisting of 270 whole-slide images. Algorithms were compared with diagnoses from human pathologists both with and without time constraints on a test data set of 129 whole-slide images (49 with and 80 without metastases). The best performing algorithms were all based on deep learning models. These outperformed pathologists when time constraints were in place, and performed comparably to pathologists without time constraints (Ehteshami Bejnordi et al. 2017).

Gulshan et al. compared the performance of the CNN vs clinicians at grading diabetic retinopathy based on retinal fundus photographs (Gulshan et al. 2016). The task was to make multiple binary predictions, some of which would determine referral for diabetic retinopathy. The baseline CNN architecture was GoogleNet v3 (Szegedy et al. 2015). The development set consisted of 128,175 images, 80% of which were used for training, and 20% for tuning. The raw data was labelled/graded by 54 US licenced ophthalmologists or

ophthalmology trainees in their final year of residency. An ensemble of ten CNNs were then trained on the same data with final prediction scores being the linear average of the CNNs that comprised the ensemble. Two validation sets consisting of 9963 and 1748 images were used to compare the CNN ensemble vs clinicians. High sensitivity and high specificity set points were used, and the algorithm performance matched that of the ophthalmologists.

Segmentation

A significant benefit from the application of ML to segmentation in imaging for diagnostics lies in the time-saving labour reduction of automation. Many valuable tools in this space are not strictly machine learning, but rather feature detection (Belevich et al. 2016). However, some of the semi and fully automated classifiers do rest upon a foundation of machine learning. The trainable Waikato environment for knowledge analysis (WEKA) segmentation is an example of a supervised learning classifier which was designed to be widely applicable to multiple data types and types of microscopy (Arganda-Carreras et al. 2017). This is distributed as a plugin for the open source FIJI imaging platform (Schindelin et al. 2012) which can interpret image data and use the WEKA data mining toolkit to perform classification given a user-trained dataset (Hall et al. 2009). The modularity as a FIJI plugin has meant that it has found widespread use amongst many types of imaging data, even tomography where substantial amounts of data can be discarded (e.g. one in 15 images are used) (Staniewicz and Midgley 2015). It has also meant that there is now widespread deployment of classifiers based on an ML toolkit that microscopists or clinicians use to augment their throughput when classifying images for diagnosis.

Machine learning has proven efficient at using automated segmentation to reduce diagnostics workload in the field histopathology in particular. Litjens et al. successfully identified micro and macro metastases of breast cancer tissue and successfully rejecting 30% of samples with only benign tissue (Litjens et al. 2016). This work relied on the application of CNNs based on widely available Theano and Pylearn toolkits, deciding on a patch size of 128 × 128 pixels and optimising over five epochs for network structure (e.g. number of layers, filters per layer, number of nodes in fully connected layers) and parameters (e.g. learning rate, momentum) with a training time of 80 and 200 min for prostate cancer and lymph nodes, respectively.

Caveats

ML applications in medicine are not without pitfalls. Cabitza et al. argue that skill reduction in medical practitioners is a distinct possibility (Cabitza et al. 2017). The quality of the output of an algorithm is also largely determined by the quality of the data, which can result in erroneous conclusions if the training set is not correctly vetted. For example, the omission of ICU admission requirements of patients with a history of asthma in the context of pneumonia led to the erroneous conclusion of better health outcomes for asthma sufferers with concurrent pneumonia compared to non-asthmatic patients with pneumonia (Caruana et al. 2015).

The output of machine learning classifiers tends to be in the form of a probability estimate between 0 and 1. Rarely do they take into account the possibility of multiple pathologies, or pathologies that may interact or augment the presentation of the other.

Diagnostic classifications themselves can also be controversial and incompletely defined. For example, overlapping conditions with different diagnostic labels could be part of a spectrum/continuum. This is possibly the reason why most clinical applications discussed above have been on binary variables such as benign vs malignant or presence vs absence. For example, inflammatory conditions hold a significant portion of clinical presentations and diagnostically have overlapping features.

The successful introduction of ML as a new diagnostic and therapeutic technique relies on it outperforming current clinical standards. While sensitivity, specificity and ROC comparisons between algorithms and clinicians on test data sets certainly add validity to algorithm performance, the gold standard for any new methodology applied in a clinical setting relies on comparable or superior performance in a randomised clinical trial.

Compliance with ethical standards

Conflicts of interest All authors declare that they have no conflicts of interest.

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

References

Arganda-Carreras I, Kaynig V, Rueden C, Eliceiri KW, Schindelin J, Cardona A, Seung HS (2017) Trainable Weka segmentation: a machine learning tool for microscopy pixel classification. Bioinformatics 33(15):2424–2426. https://doi.org/10.1093/bioinformatics/btx180

Belevich I, Joensuu M, Kumar D, Vihinen H, Jokitalo E (2016) Microscopy image browser: a platform for segmentation and analysis of multidimensional datasets. PLoS Biol 14(1):1–13. https://doi.org/10.1371/journal.pbio.1002340

Bottou L (2010) Large-scale machine learning with stochastic gradient descent. Proceedings of COMPSTAT'2010, 177–186. doi: https://doi.org/10.1007/978-3-7908-2604-3_16

Breiman L (1996) Bagging predictors. Mach Learn 24(2):123–140. https://doi.org/10.1007/BF00058655

Cabitza F, Rasoini R, Gensini GF (2017) Unintended consequences of machine learning in medicine. JAMA. https://doi.org/10.1001/jama.2017.7797

Caruana R, Lou Y, Gehrke J, Koch P, Sturm M, Elhadad N (2015) Intelligible models for healthcare. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD'15 (pp. 1721–1730). doi: https://doi.org/10.1145/2783258.2788613

Chartrand G, Cheng PM, Vorontsov E, Drozdzal M, Turcotte S, Pal CJ, … Tang A (2017) Deep learning: a primer for radiologists. RadioGraphics 37(7):2113–2131.https://doi.org/10.1148/rg.2017170077

Deng J, Dong W, Socher R, Li L-J, Li K, Li F-F (2009) ImageNet: a large-scale hierarchical image database. In 2009 IEEE Conference on Computer Vision and Pattern Recognition (pp. 248–255). doi https://doi.org/10.1109/CVPRW.2009.5206848

Ehteshami Bejnordi B, Veta M, Johannes van Diest P, van Ginneken B, Karssemeijer N, Litjens G, … Venâncio R (2017) Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. JAMA 318(22):2199. doi https://doi.org/10.1001/jama.2017.14585

Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, Thrun S (2017) Dermatologist-level classification of skin cancer with deep neural networks. Nature 542(7639):115–118. https://doi.org/10.1038/nature21056

Fawcett T (2006) An introduction to ROC analysis. Pattern Recogn Lett 27(8):861–874. https://doi.org/10.1016/j.patrec.2005.10.010

Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, … Webster DR (2016) Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA 316(22):2402–2410. doi:https://doi.org/10.1001/jama.2016.17216

Hall MA, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH (2009) The WEKA data mining software: an update. SIGKDD Explor Newsl 11(1):10–18. https://doi.org/10.1145/1656274.1656278

Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning. Springer New York, New York

Kohavi R (1995) A study of cross-validation and bootstrap for accuracy estimation and model selection. In Appears in the International Joint Conference on Articial Intelligence (IJCAI), pp. 1–7. doi https://doi.org/10.1067/mod.2000.109031

Krizhevsky A, Sutskever I, Hinton GE (2012) ImageNet classification with deep convolutional neural networks. Adv Neural Inf Proces Syst:1–9. https://doi.org/10.1016/j.protcy.2014.09.007

LeCun Y (1988) A theoretical framework for back-propagation. Proceedings of the 1988 connectionist models summer school. doi https://doi.org/10.1007/978-3-642-35289-8

Litjens G, Sánchez CI, Timofeeva N, Hermsen M, Nagtegaal I, Kovacs I, … Van Der Laak J (2016) Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. Sci Rep 6: 1–11. doi https://doi.org/10.1038/srep26286

Litjens G, Kooi T, Bejnordi BE, Arindra A, Setio A, Ciompi F et al (2017) A survey on deep learning in medical image analysis. Med Image Anal 42:60–88. https://doi.org/10.1016/j.media.2017.07.005

Marchetti, M. A., Codella, N. C. F., Dusza, S. W., Gutman, D. A., Helba, B., Kalloo, A., … Halpern, A. C. (2018). Results of the 2016 International Skin Imaging Collaboration International Symposium on Biomedical Imaging challenge: comparison of the accuracy of computer algorithms to dermatologists for the diagnosis of melanoma from dermoscopic images. doi https://doi.org/10.1016/j.jaad.2017.08.016

Pan SJ, Yang Q (2010) A survey on transfer learning. IEEE Trans Knowl Data Eng. https://doi.org/10.1109/TKDE.2009.191

Parker DB (1985) Learning-logic: casting the cortex of the human brain in silicon Technical report Tr-47, Centre for computational research in economics and management science. MIT, Cambridge

Rumelhart DE, Hinton GE, Williams RJ (1986) Learning representations by back-propagating errors. Nature 323(6088):533–536. https://doi.org/10.1038/323533a0

Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S et al (2015) ImageNet large scale visual recognition challenge. Int J Comput Vis 115(3):211–252. https://doi.org/10.1007/s11263-015-0816-y

Schindelin, J et al (2012) Fiji: an open-source platform for biological-image analysis, Nature methods 9(7):676–682

Silver D et al (2017) Mastering the game of go without human knowledge. Nature 550(7676):354–359

Spontón H, Cardelino J (2015) A review of classic edge detectors. IPOL 5:90–123. https://doi.org/10.5201/ipol.2015.35

Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R (2014) Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res 15:1929–1958. https://doi.org/10.1214/12-AOS1000

Staniewicz L, Midgley PA (2015) Machine learning as a tool for classifying electron tomographic reconstructions. Adv Struct Chem Imaging 1(1):9. https://doi.org/10.1186/s40679-015-0010-x

Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z (2015) Rethinking the inception architecture for computer vision. https://doi.org/10.1109/CVPR.2016.308