



Demystification of AI-driven medical image interpretation: past, present and future

Peter Savadjiev¹ · Jaron Chong^{1,2} · Anthony Dohan^{1,2,3} · Maria Vakalopoulou^{4,5} · Caroline Reinhold^{1,2} · Nikos Paragios^{4,6} · Benoit Gallix^{1,2}

Received: 15 June 2018 / Revised: 5 July 2018 / Accepted: 17 July 2018 / Published online: 13 August 2018
© European Society of Radiology 2018

Abstract

The recent explosion of ‘big data’ has ushered in a new era of artificial intelligence (AI) algorithms in every sphere of technological activity, including medicine, and in particular radiology. However, the recent success of AI in certain flagship applications has, to some extent, masked decades-long advances in computational technology development for medical image analysis. In this article, we provide an overview of the history of AI methods for radiological image analysis in order to provide a context for the latest developments. We review the functioning, strengths and limitations of more classical methods as well as of the more recent deep learning techniques. We discuss the unique characteristics of medical data and medical science that set medicine apart from other technological domains in order to highlight not only the potential of AI in radiology but also the very real and often overlooked constraints that may limit the applicability of certain AI methods. Finally, we provide a comprehensive perspective on the potential impact of AI on radiology and on how to evaluate it not only from a technical point of view but also from a clinical one, so that patients can ultimately benefit from it.

Key Points

- *Artificial intelligence (AI) research in medical imaging has a long history*
- *The functioning, strengths and limitations of more classical AI methods is reviewed, together with that of more recent deep learning methods.*
- *A perspective is provided on the potential impact of AI on radiology and on its evaluation from both technical and clinical points of view.*

Keywords Diagnostic imaging · Artificial intelligence (AI) · Machine learning · Computer-assisted image processing · Computer-assisted image interpretation

✉ Benoit Gallix
benoit.gallix@mcgill.ca

¹ Department of Diagnostic Radiology, McGill University, Montreal, QC, Canada

² Department of Diagnostic Radiology, McGill University Health Centre, Montreal, QC, Canada

³ Department of Body & Interventional Imaging, Hôpital Lariboisière-AP-HP, Université Diderot-Paris 7 and INSERM U965, 75475 Paris Cedex 10, France

⁴ Ecole CentraleSupélec, 91190 Gif-sur-Yvette, France

⁵ Inria Saclay/Ile-de-France, 91120 Palaiseau, France

⁶ TheraPanacea, 75014 Paris, France

Abbreviations

AI	Artificial intelligence
ANN	Artificial neural network
LASSO	Least absolute shrinkage and selection operator
PICO	Patient/problem, intervention, comparison intervention and outcomes
TB	Tuberculosis

Introduction

Taking advantage of the increasingly large amount of labelled digital information available in every area of technological activity, artificial intelligence (AI), and in particular methods such as deep learning, have recently achieved impressive

results with ‘big data’ in many different domains [1]. The healthcare sector, rich in complex data and processes, is also being transformed, and radiology will be one of the first specialties to be affected. Despite some opinions voiced in the public space, AI is unlikely to replace human radiology expertise. However, in the long term, AI may augment existing computer-based tools or even partially replace human medical expertise for certain specialised and/or repetitive tasks, such as computing targets in radiotherapy, detecting disease indicators in images or measuring longitudinal disease burden [2].

At the same time, technological solutions for domains rich in labelled data do not necessarily translate easily to the healthcare sector. In order to recognise general object categories in images, popular internet platforms train deep learning algorithms on hundreds of millions of labelled data easily accessible in large image databases [e.g. the ImageNet project, <http://image-net.org>]. These data sets usually consist of 2D images with much fewer pixels than typical 2D or 3D medical images. Furthermore, such images do not come with regulatory or ethical constraints, the correct solution to whether an image shows (for instance) a cat, a dog or a car does not require an expert’s annotation with strong inter-user agreement, and the cost of failure is low. In contrast, creating a gold standard of medical diagnosis is often challenging and sometimes unclear. Medical data is subject to multiple privacy and regulatory considerations and is not easily accessible. These accessibility constraints can make it very difficult to curate high-quality image data sets.

Methods such as deep learning, however, are only the latest in a long history of AI-based approaches to medical image analysis that span several decades. This effort has produced advanced methods for the automation of many tasks in radiological image processing [e.g. 3]. These tasks include organ localisation; organ segmentation; the detection and segmentation of lesions within the organ; registration for the longitudinal follow-up of disease progress, or for combining image data of different modalities; and finally, the analysis of image patterns within the organ and the tumour region, with the goal of revealing associations with clinical outcome. We provide a review on how the computational technology required to achieve these objectives evolved during the past decades in order to provide a context for the latest developments. We review the functioning and limitations of these methods and present a perspective on the relationship between AI and the traditional discovery process in medicine.

Symbolic interpretation of medical images

The first attempts to develop automatic interpretation of medical images were based on human decisional models, performing high-level (i.e. symbolic) image interpretation. In

the 1980s, AI approaches to medical imaging relied on so-called expert systems that performed logical inference based on rules derived from a human approach to image processing [e.g. 4]. In practice, this involved image processing operations that would be considered simple by modern standards, for instance binarising/thresholding an image and looking for geometric structures such as lines, circles or trapezoids. A set of logical rules would then deduce image content based on the presence/absence of such structures [e.g. 5].

The main strength of such methods is their ability to provide solutions that are fully human interpretable and therefore easily acceptable by physicians. In addition, these models have excellent generalisation since decisions are taken on the basis of human medical knowledge. On the negative side, the mere encoding of human decisional processes did not prove to add value to medical practice, especially given the limited computational resources of that era. While expert systems were successful in some other domains, such as business analysis or manufacturing, they did not produce a significant impact in medicine. Table 1 provides a summary for this type of approach as well as for the more recent classes of methods discussed next.

Probabilistic interpretation of medical images

The second generation of algorithms was also inspired by the human decisional chain, but instead of adopting a strongly symbolic interpretation of observations, it moved towards a statistical/probabilistic one. This category of methods is also based on a model derived from human expertise; however, the model’s parameters are computed from a labelled reference data set using probabilistic methods that determine the most likely solutions [e.g. 6–9]. In this manner, a statistical model—sometimes called an atlas—is created to represent the inherent signal variability in the reference population (Fig. 1).

Image/organ segmentation—one of the pillars of medical imaging—is a classic example to demonstrate this concept. A structure of interest in a new image can be described by matching its appearance to the population atlas. If the atlas is constructed from a healthy population of images, then lesions in the new image can be modelled as abnormal outliers relative to the expected appearance of healthy tissues (Fig. 2). This approach is commonly used in neuroimaging for modelling brain anatomical structure or for lesions arising from traumatic injuries, cancer, multiple sclerosis or other diseases [e.g. 10–12].

Such approaches inherit numerous strengths but also important limitations. On the positive side, these methods aggregate information across populations and experts, couple them with a human-understandable statistical model and therefore are endowed with good acceptability and good generalisation properties. However, the choice of the statistical model and its

Table 1 Summary of the different types of approaches to AI in medical imaging

Generation of algorithms	Prominent example	Popularity era	Overall principle	Example application: segmentation	Strengths and weaknesses
Symbolic	Expert systems	1980s–1990s	Encode human knowledge as a set of rules	A set of heuristic rules describe boundary detection and region merging operations	<ul style="list-style-type: none"> Fully human interpretable, good generalisation Performs worse than human expertise
Probabilistic	Atlas-based models	1990s–present	A statistical model is designed from human knowledge, but its parameters are inferred from data	Abnormal regions are modelled as outliers relative to an atlas of normal population variability. See Figs. 1 and 2	<ul style="list-style-type: none"> Combines human-understandable model with information across populations Not always clear what is the most appropriate model for a given problem
Model-free/ data-driven	Deep learning	2010s–present	Minimal human knowledge involvement outside of supervised training stage. New knowledge is discovered by automated data mining	A classification of pixels is assigned on the basis of training on large sets of presegmented images	<ul style="list-style-type: none"> Exceeds human performance and that of other methods in integrating information across large data sets Requires large amounts of processing power, typically does not generalise well, and properly qualified large databases are not easily available

ability to cope with the observed anatomical variability play a significant role in the overall performance. Furthermore, in order to build an appropriate model and then successfully

apply it, all images need to be transformed to the same reference space via deformable registration, which is a highly challenging problem on its own [13].

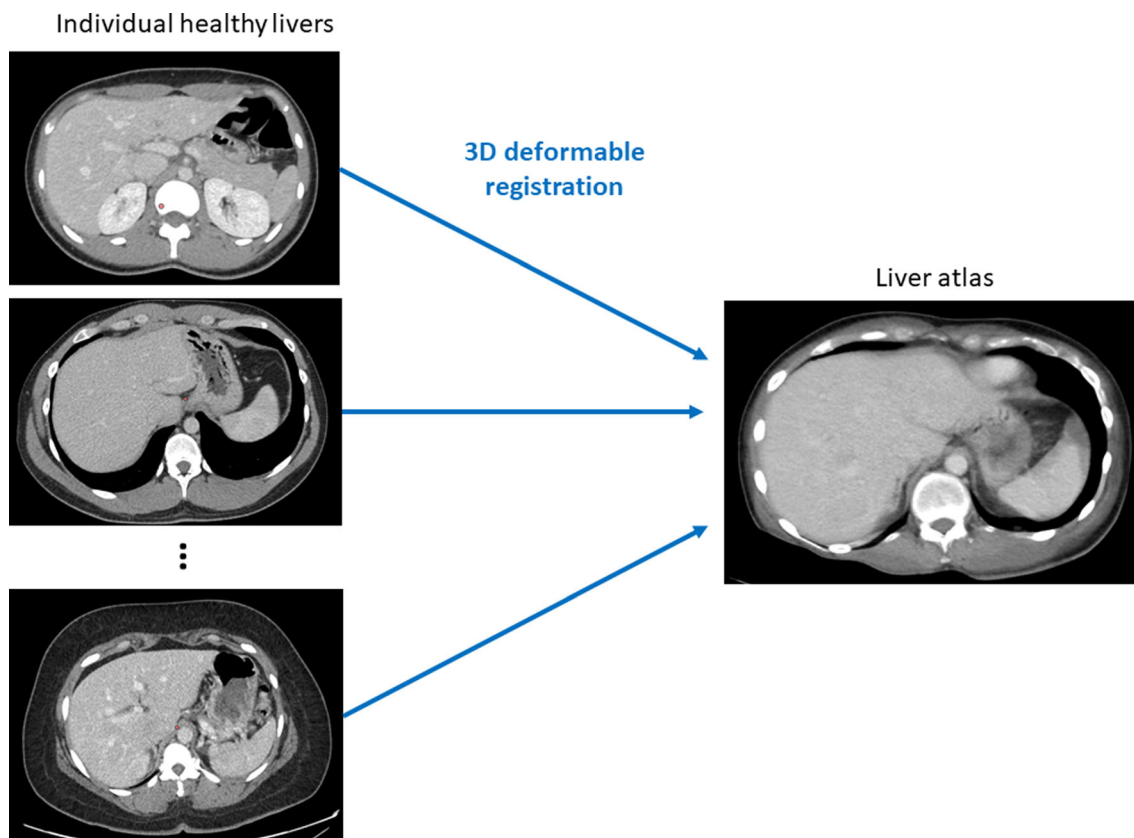


Fig. 1 An atlas encoding the variability of the appearance of healthy livers in a reference population can be created via deformable registration of individual liver images into a common target space. In this example, the registration was obtained with the method of [28]

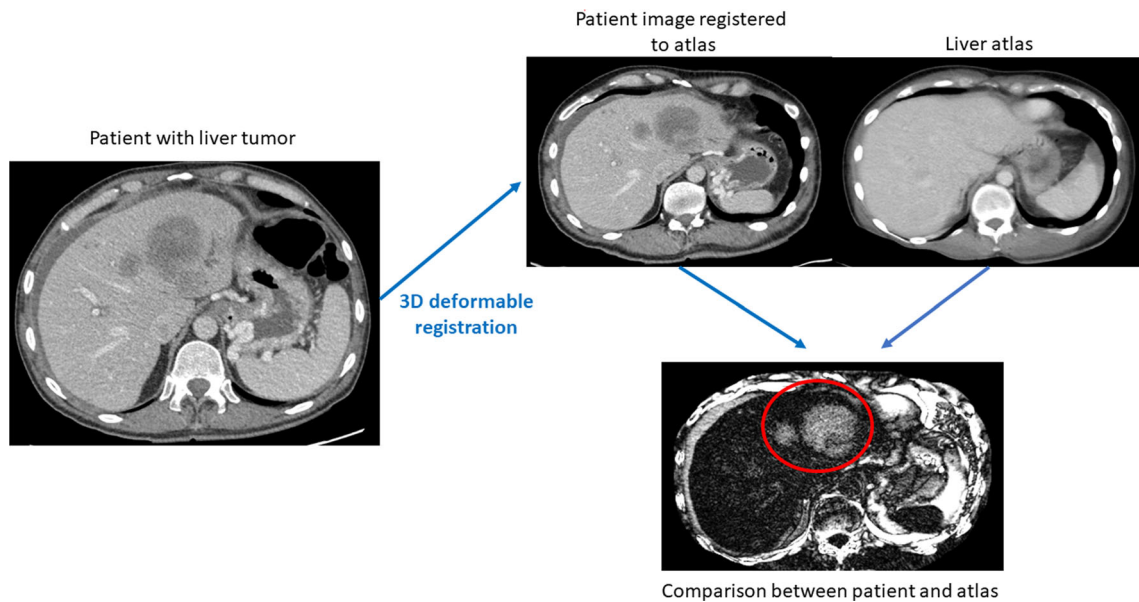


Fig. 2 When an individual image containing a lesion is compared to a healthy population atlas, the lesion will appear as an abnormal outlier region. In this manner, lesions can be detected and mapped without the need for explicit segmentation

Data-driven/model-free approaches to automated knowledge discovery

A limitation of methods that rely on explicit expert knowledge is that transforming human-level expertise into suitable computational models can be challenging, especially when human knowledge is incomplete. Therefore, there has recently been a surge of interest in a different kind of approach—the ‘discovery science’ approach, which emphasises the mining of large amounts of data to discover new patterns and help formulate new hypotheses [e.g. 14]. As such, these approaches are model-free and purely data-driven. A currently popular application of these methods is known as ‘texture analysis’ or

‘radiomics’, where features computed over a segmented tumour region are used to classify different tumour types [15].

In the context of radiology, methods in this category learn a characteristic representation of the appearance of organs and tissues from a set of training images. This approach is referred to as supervised learning when the training images have previously been (manually) labelled. If there is no such labelling, methods exist to automatically detect groupings in the data with what is known as unsupervised learning. In both cases, learning is based on a large set of basic image descriptors (features) that are automatically extracted from the image. The goal is to use these features to discover a separation between classes in high-dimensional feature spaces; see Fig. 3 for a simple two-

Fig. 3 Example of two-dimensional decision boundaries between two groups of data points, with increasing complexity from **a** to **b** to **c**; **a** is likely to be too simple, whereas **c** is likely to be overfitted

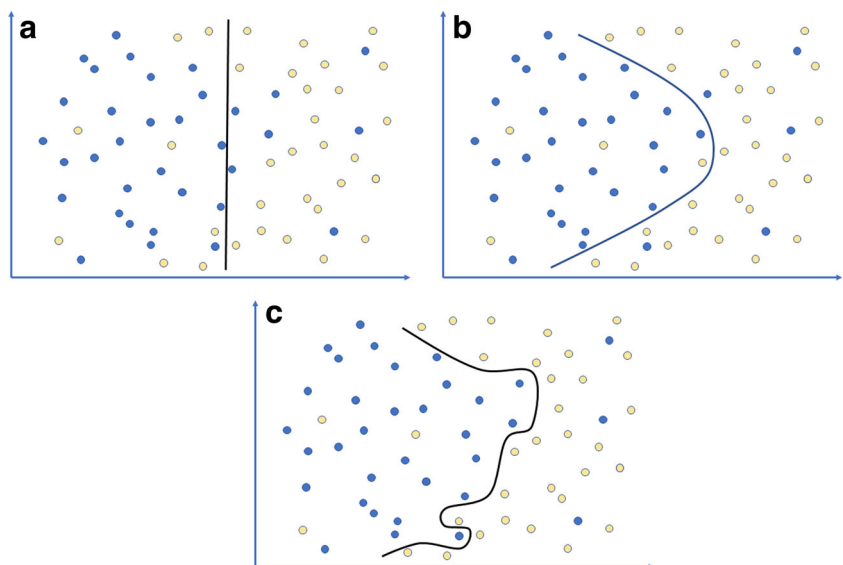
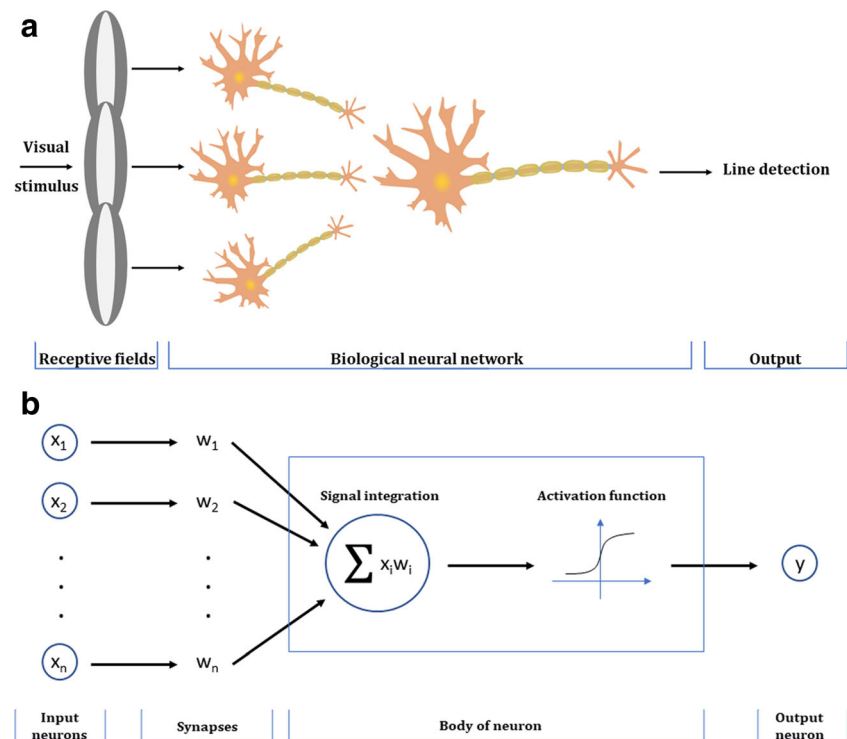


Fig. 4 **a** Schematic representation of a simple biological neural network. **b** Analogous artificial neural network. Adapted from [21]



dimensional example. Different statistical and machine learning techniques exist to do this, for instance logistic regression, support vector machines [16], decision trees [17] and many others.

However, when the number of features computed is much larger than the number of samples in a typical data set, the risk of overfitting is increased. Overfitting occurs when the system learns idiosyncratic properties of the training data, i.e. properties that are too specific to the training data and do not generalise well. This will prevent the system from reliably fitting new data or predicting future observations in a reliable manner. To reduce the risk of overfitting, one option is to perform feature selection, i.e. reducing the number of features under consideration by keeping only the most relevant features and discarding the rest [e.g. 18, 19]. One popular feature selection method is the least absolute shrinkage and selection operator (LASSO) [20], which alters standard regression methods by selecting only a subset of the available covariates, as opposed to using all of them. In doing so, it removes predictor variables that may contribute to overfitting. As another example of feature selection, in experimental paradigms that involve human interaction, for instance the manual segmentation of lesions, it is possible to determine the reproducibility of features across different human operators and then discard those features that are not reproducible.

As an example of potential overfitting, on a small training set of images of different tumour grades, given a sufficiently large number of features, one could in principle always find a particular model that provides perfect classification of the training images into different tumour grades. However, such

a model is not likely to have a practical value unless it is validated on an independent data set. In fact, it is customary in machine learning and statistical approaches to classification to work with three separate data sets. The first is the training set, used to optimise model parameters. The second data set—the validation data set—is then used to assess the quality of the training and the potential presence of overfitting. Finally, the third data set, called the test data set, is used to provide a final, unbiased report of the method's performance.

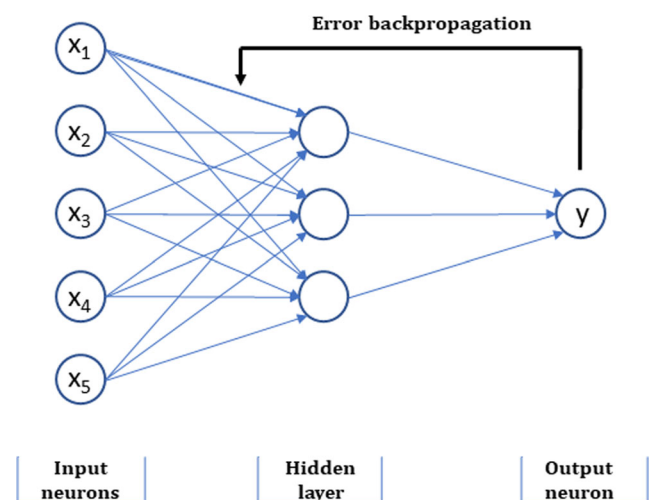


Fig. 5 Schematic representation of an artificial neural network with backpropagation of error signals during the training process. At a very abstract level, backpropagation is analogous to human training where experiences improve a doctor's skills and knowledge

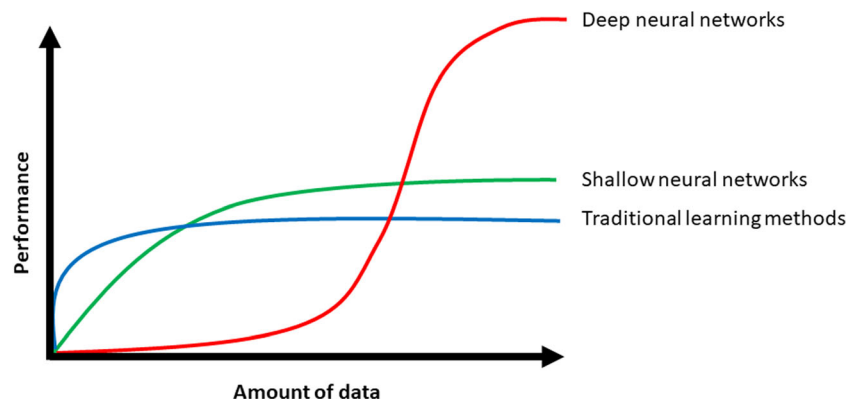


Fig. 6 Schematic representation of the impact of data availability on the performance of various machine learning methods. Traditional machine learning methods, using human-crafted features, perform best when only a small amount of data is available. Increasingly deeper neural networks

(i.e. with a larger number of layers) perform increasingly better as the amount of available data is increased. Adapted from Easy Solutions Inc. [<http://blog.easysol.net/building-ai-applications/>]

The advantage of data-driven approaches is that they do not require expert modelling of the specific problem at hand. Instead, the underlying assumption is that provided a large enough training set, the meaningful information and variability in the data can be represented as a function of a large set of basic image features. On the negative side, these methods are often myopic (decisions are taken at a local scale where global context is ignored), not modular (for each task, user-driven adjustment of the feature space might be required) and their generalisation needs to be verified with independent validation sets. In addition, methods such as radiomics or texture analysis suffer from a lack of standardisation, since many of the features they work with depend on specific image acquisition parameters that can vary according to scanner hardware and software.

Deep learning/sub-symbolic artificial intelligence

Typical data-driven approaches in radiology work with features designed to reflect properties of the data seen as important from a human radiologist's point of view, such as density, heterogeneity of tumours, tumour shape, etc. However, a class of machine learning methods have recently gained popularity because they can automatically discover the best features for a given task, without requiring human 'feature engineering'. This class of methods is known as deep learning and can oftentimes exceed human performance for specific tasks [1, 21].

Architecture

Deep learning systems are built upon an architecture of artificial neural networks (ANNs). Each ANN consists of an array of processing units, known as artificial 'neurons' or nodes. Each such node takes as input a set of feature values, multiplied by a corresponding weight. The node then sums up the weighted evidence it receives as input, and then passes it through a nonlinear activation function, which determines

the neuron's output (Fig. 4). Essentially, each artificial neuron makes a decision based on weighted evidence. This design is intended to mimic, in a simplified manner, the behaviour of a biological neuron, which integrates multiple inputs from synapses on its dendrites and sends an output signal down its axon, as input to another neuron.

Based on this paradigm, hierarchical networks of neurons can be used to encode very complex nonlinear functions. Deep learning systems implement this concept with multilayer hierarchies of ANNs. The first layer in the hierarchy contains nodes that are directly sensitive to the input data, such as individual pixel values in an image. The last layer represents output values such as classification results. The intermediate layers compute intermediate feature representations that become more abstract as the information travels towards deeper layers. Ultimately, the deep learning network is designed to convert raw activation signals from the input layer to target solution values determined from the activation pattern in the last layer.

Training and learning

Deep learning networks in use today consist of thousands of nodes and millions of connections, resulting in millions of network parameters. For the network to be useful, all these parameters need to be adjusted through a training process. Each element of the training set (e.g. each image) is given as input to the network, and the network's response based on its current configuration is recorded. Following this, the network's current output prediction is quantitatively compared with the true prediction through an error function, termed a 'loss function'. Through a process known as back-propagation [1, 22], each parameter of the network is adjusted by a small increment in a direction to minimise the network's loss function. This process is repeated iteratively, until the network reaches a state minimising the loss function (Fig. 5).

Table 2 The PICO approach adapted to developing focused clinical questions for AI-based research in medical Imaging

Framework	Key step	Particularity for AI in diagnostic radiology	Example: AI for automatic reading of chest radiography in the native population of Canada for TB diagnosis
Population (or patient problem)	Describe the patient population. What are the most important patient characteristics and/or the problem type?	Describe the specific disease or condition	Population at high risk of tuberculosis (TB) in the Canadian northern territories
Intervention	Describe the diagnostic test(s) being considered	Describe: - Method of data acquisition - Process of data quality assurance - Method of building the image database for this patient group	All chest radiographies performed in the territory will be prospectively archived in a dedicated server Specific training for technologists will be put in place to ensure a constant quality of acquisitions
Comparison or control	Describe the current “gold standard” test for that disease or condition	Comparison of AI to human performance (radiologist reading)	Performance of AI automatic analysis of chest radiographs will be compared to that of expert and non-expert radiologists
Outcome	Determine a clinical outcome that will be used to evaluate the performance of the intervention For instance, measures of classification performance, e.g. sensitivity, specificity, accuracy	Determine how AI will be applied to screening, triage, staging Describe how the data set will be annotated for AI training	AI will be used as a triage tool for TB. The diagnosis of TB will be based on subsequent clinical biological and other imaging modality examination. The final diagnosis will be based on an expert panel

This training process can be very computationally intense for networks with a large number of layers and nodes, and this is what has kept artificial neural networks in relative obscurity for many decades. Research on neural networks in AI dates back to the 1950s [23]. However, the modern resurgence in neural networks owes much to two recent developments that make training such networks practical and tractable: (1) the arrival of inexpensive and powerful computing hardware and memory, and (2) the availability of large quantities of labelled digital data that can be used for the training process. These two factors are crucial in the modern success of deep learning algorithms.

Limitations of deep learning

Despite its success and popularity in a wide variety of tasks, deep learning comes with several limitations.

Inefficient use of data and processing power

To learn a particular task, a deep learning algorithm requires several orders of magnitude more data points than other types of learning algorithms, let alone a human. When such amounts of data are available, then deep learning can outperform any other method. However, more traditional methods tend to perform significantly better with smaller amounts of data, as illustrated schematically in Fig. 6. The deep learning community has recognised this problem and is now turning towards developing the next generation of algorithms, which will be much more complex mathematically, but would present a more efficient use of data; see, for instance, [24] for a layman’s introduction to this topic.

Limited generalisation and model fragility

A given deep neural network can typically be applied only to a single narrowly defined task, and its performance is tied to a particular data set and label set. In the case of radiological applications, if deep learning is to be used to classify tumour images, different deep networks would need to be retrained and relabelled for every type of organ, tumour, modality and orientation. In adversarial situations, even minor changes to the input data, often invisible to the human eye, can result in dramatically different classifications [e.g. 25]. This is why, even if deep learning were to be successfully integrated with radiology, human verification will most likely be required.

A medical approach to AI research in radiology

The current AI trends centred around deep learning have produced a pressure to shift from hypothesis-driven to data-

driven research. The data-driven approach can be powerful and lead to novel insights; however, it cannot replace the cognitive integration of complex information combining the semeiological analysis of images features in a precise anatomical and physiopathological context. These methodological approaches have been refined over centuries of scientific and medical thought, and they allow for solutions that are still inaccessible to AI.

AI brings new opportunities, but the fundamental principles of clinical reality do not change. In order to effect real impact on patient care, AI-based research in medical imaging must adhere to first principles in medical science. Research hypotheses, whether AI-based or not, must be clinically relevant and answerable in the clinical domain. In Table 2 we have adapted the classical PICO approach [26] to develop focused clinical questions adapted to AI-based research in medical Imaging. PICO is an acronym for patient/problem, intervention, comparison intervention and outcomes.

Because of a certain invincible aura that AI has in popular perception, a common mistake is to expect AI to find analysable information in medical images even where none exists or for which there is no biological plausibility. Akin to perennial concerns in medical research of data dredging and p-hacking [e.g. 27], improperly designed AI experiments could misinform, mislead or without critical analysis could result in patient harm.

In the current literature, AI-based methods are usually evaluated by comparison with expert individual or consensus radiologist performance. However, the influence of clinical history and context, differences in population incidence, and the inherent subjectivity of interpretations can make such experimental paradigms unreliable. Rather, training labels where possible should relate to meaningful patient outcomes and gold standards that exist beyond radiology. At the same time, it is important to compare human and AI performance in terms of efficacy, reproducibility, turnaround time and cost. In that sense, AI should be compared not only to the best experts in the field but also to general practitioners, and possibly also to healthcare systems with limited human resources, for instance those in developing countries. In certain high-need settings, AI may bring additional value, in particular as a screening tool.

Conclusion

Novel computational tools derived from AI are likely to transform radiological practice. The greatest challenge at this time—and also the most exciting opportunity—lies in determining which particular clinical tasks in radiology are the most and least likely to benefit from AI algorithms, given the power but also the limitations of such algorithms. At the same time, regardless of the technology used, it is important to

keep in mind that proper clinical hypotheses remain primordial, and an appropriate validation of AI-based methods against actual clinical outcome measures of patient well-being remains the most important measure of success.

Funding The authors state that this work has not received any funding.

Compliance with Ethical Standards

Guarantor The scientific guarantor of this publication is Dr. Benoit Gallix.

Conflict of Interest Professor Nikos Paragions declares a relationship with the following company: TheraPanacea, Paris, France.

The other co-authors of this manuscript declare no relationships with any companies whose products or services may be related to the subject matter of the article.

Statistics and Biometry No complex statistical methods were necessary for this paper.

Informed Consent Written informed consent was not required for this study because this is a review article, no study was performed.

Ethical Approval Institutional review board approval was not required because this is a review article and no study was performed.

References

1. LeCun Y, Bengio Y, Hinton G (2015) Deep learning. *Nature* 521: 436–444
2. Tang A, Tam R, Cadrin-Chênevert A et al Canadian Association of Radiologists (CAR) Artificial Intelligence Working Group (2018) Canadian Association of Radiologists white paper on artificial intelligence in radiology. *Can Assoc Radiol J* 69:120–135
3. Summers RM (2016) Progress in fully automated abdominal CT interpretation. *AJR Am J Roentgenol* 207:67–79
4. Matsuyama T (1989) Expert systems for image processing: knowledge-based composition of image analysis processes. *Comput Vision Graph* 48:22–49
5. Stansfield SA (1986) ANGY: a rule-based expert system for automatic segmentation of coronary vessels from digital subtracted angiograms. *IEEE Trans Pattern Anal Mach Intell* 2:188–199
6. Park H, Bland PH, Meyer CR (2003) Construction of an abdominal probabilistic atlas and its application in segmentation. *IEEE Trans Med Imaging* 22:483–492
7. Warfield SK, Zou KH, Wells WM. (2004) Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. *IEEE Trans Med Imaging* 23:903–21
8. Okada T, Linguraru MG, Hori M, Summers RM, Tomiyama N, Sato Y (2015) Abdominal multi-organ segmentation from CT images using conditional shape-location and unsupervised intensity priors. *Med Image Anal* 26:1–18
9. Iglesias JE, Sabuncu MR (2015) Multi-atlas segmentation of biomedical images: a survey. *Med Image Anal* 24:205–219
10. Van Leemput K, Maes F, Vandermeulen D, Colchester A, Suetens P (2001) Automated segmentation of multiple sclerosis lesions by model outlier detection. *IEEE Trans Med Imaging* 20:677–688

11. Prastawa M, Bullitt E, Moon N, Van Leemput K, Gerig G (2003) Automatic brain tumor segmentation by subject specific modification of atlas priors. *Acad Radiol* 10:1341–1348
12. Erus G, Zacharaki EI, Davatzikos C (2014) Individualized statistical learning from medical image databases: application to identification of brain lesions. *Med Image Anal* 18:542–554
13. Viergever MA, Maintz JBA, Klein S, Murphy K, Staring M, Pluim JPW (2016) A survey of medical image registration - under review. *Med Image Anal* 33:140–144
14. Kraus WL (2015) Editorial: would you like a hypothesis with those data? Omics and the age of discovery science. *Mol Endocrinol* 29:1531–1534
15. Aerts HJ (2016) The potential of radiomic-based phenotyping in precision medicine: a review. *JAMA Oncol* 2:1636–1642
16. Cortes C, Vapnik V (1995) Support-vector networks. *Mach Learn* 20:273–297
17. Quinlan JR (1986) Induction of decision trees. *Mach Learn* 1:81–106
18. Matzner-Lober E, Suehs CM, Dohan A, Molinari N (2018) Thoughts on entering correlated imaging variables into a multivariable model: application to radiomics and texture analysis. *Diagn Interv Imaging* 99:269–270
19. Guyon I, Elisseeff A (2003) An introduction to variable and feature selection. *J Mach Learn Res* 3:1157–1182
20. Tibshirani R (1996) Regression shrinkage and selection via the lasso. *J R Stat Soc B* 58:267–288
21. Chartrand G, Cheng PM, Vorontsov E et al (2017) Deep learning: a primer for radiologists. *Radiographics* 37:2113–2131
22. Werbos P (1974) Beyond regression: new tools for prediction and analysis in the behavioral sciences. PhD thesis, Harvard Univ
23. Rosenblatt F (1957). The Perceptron—a perceiving and recognizing automaton. Report 85-460-1, Cornell Aeronautical Laboratory
24. Lawrence N (2016) Deep learning, Pachinko and James Watt: efficiency is the driver of uncertainty. <http://inverseprobability.com/2016/03/04/deep-learning-and-uncertainty>. Accessed 23 May 2018
25. Szegedy C, Zaremba W, Sutskever I et al (2013) Intriguing properties of neural networks. arXiv:1312.6199
26. Richardson WS, Wilson MC, Nishikawa J, Hayward RS (1995) The well-built clinical question: a key to evidence-based decisions. *ACP J Club* 123:A12–A13
27. Simmons JP, Nelson LD, Simonsohn U (2011) False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychol Sci* 22:1359–1366
28. Ferrante E, Dokania PK, Marini R, Paragios N (2017) Deformable registration through learning of context-specific metric aggregation. *Machine Learning in Medical Imaging Workshop. MLMI (MICCAI 2017)*, Sep 2017, Quebec City, Canada