Name of Journal: *Artificial Intelligence in Cancer*

Manuscript NO: 75772

Manuscript Type: MINIREVIEWS

Building and evaluating an AI algorithm: A practical guide for practicing oncologists

Anupama Ramachandran, Deeksha Bhalla, Krithika Rangarajan, Raja Pramanik, Subhashis Banerjee, Chetan Arora

Abstract

The use of machine learning and deep learning has enabled many applications, previously thought of as being impossible. Among all medical fields, cancer care is arguably the most significantly impacted, with precision medicine now truly being a possibility. The effect of these technologies, loosely known as Artificial Intelligence (AI) is particularly striking in fields involving images (such as radiology and pathology) and fields involving large amount of data (such as genomics). Practicing oncologists are often confronted with new technologies, claiming to predict response to therapy or predict the genomic make-up of patients. Understanding these new claims and technologies requires a deep understanding of the field. In this review, we provide an overview of the basis of deep learning. We describe various common tasks and their data requirements so that oncologists are equipped to start such projects, as well as evaluate algorithms presented to them.

INTRODUCTION

Artificial Intelligence (AI) has touched many areas of our everyday life. In medical practice also, it has shown great potential in several studies [1]. The implications of use of AI in oncology are profound, with applications ranging from assisting early screening of cancer to personalization of cancer therapy. As we enter this exciting transformation, practicing oncologists in any sub-field of oncology are oftentimes faced with various studies and products claiming to achieve certain results. Verifying these claims and implementing these in clinical practice remains an uphill task.

This is an educational review, through which we will attempt to familiarize the reader with AI technology in current use. We first explain some basic concepts, in order to understand the meaning of techniques labelled as AI, then move into explaining the various tasks that can be performed by AI. In each we provide information on what kind of data would be required, what kind of effort would be required to annotate these images, as well as how to assess networks based on these tasks, for the benefit of those oncologists wishing to foray into the field for research, or for those wishing to implement these algorithms in their clinical practice.

WHAT IS ARTIFICIAL INTELLIGENCE: BASIC PRINCIPLES

AI refers to a broad, non-specific term referring only to the "intelligence" in a specific task performed, irrespective of the method used. Machine learning (ML) is a subgroup of AI, and deep learning (DL) is a further sub-group of ML, which are data-driven approaches. Unlike traditional software engineering where a set of rules is defined upon which the computer's outputs are based, ML involves learning of rules by "experience" without "explicit programming" [2,3]. What this means, is given a lot of data which includes a set of inputs, and the ideal outputs (training data), the task of machine learning is to understand a pattern within this set of inputs which result in outputs closest to the ideal output. (**Figure 1**). The process of training the model is explained in **Figure 2**.

To understand this in medical terms, say the task of an AI system is to predict the survival of patients, given the stage of a particular tumour. If we were to use traditional software engineering, we would have to feed the median survival for each stage into the model, and teach the model to output the number corresponding to a particular stage. Whereas in case of a machine learning model, we would simply give as input, the stage and survival information of a few thousand patients. The model would learn the rules involved in making this prediction. While in the former case, we defined the rules (that is if stage= X then survival= Y), in the latter we only provided data, and the ML model deciphered the rules. While the former is rigid, that is, if a new therapy alters the survival, we would have to change the rules to accommodate the change, the latter learns with experience. As new data emerges, the ML model would learn to update the rules such that it can dynamically make accurate predictions. In addition, the machine learning model can take multiple inputs, say level of tumour markers, age, general condition, blood parameters into account, in addition to the stage of the patient, and personalise the survival prediction of a particular patient.

The above example also illustrates why ML models are data intensive. A good model needs to see a lot of data, with adequate variability in parameters to make accurate predictions. For the same reason, AI has bloomed in disciplines which have a lot of digital data available, this includes ophthalmology, dermatology, pathology, radiology and genomics. However as curated digital data emerges in all fields it is likely to touch and transform all fields of medical practice.

WHAT ARE NEURAL NETWORKS?

A particular kind of machine learning algorithm, called neural network, has been particularly effective in performing complex tasks. A neural network takes inspiration from a biological neuron, where it receives several inputs, performs a certain calculation, goes through an activation function, where similar to a biological neuron, a decision on whether it should fire or not is made. When there are a number of layers of these mathematical functions, the network is known as a "Deep Neural Network" (DNN), and the process is called "Deep Learning" (DL) (**Figure 3).** A deep neural network is capable of handling a large amount of data, and defining complex functions, which explains its ability to perform complex classification tasks and predictions.

A specific kind of DL, called Convolutional Neural Networks (CNNs) have performed particularly well in image-related tasks. CNNs use "filters" which are applied to images, similar to the traditional image processing techniques. "Convolving" with a filter (a mathematical operation) results in highlighting certain features of an image.

Given an input set of images, a CNN basically learns what set of filters highlights features of a particular image, most relevant to a given task. In other words, a CNN is learning the features of an image that may be crucial to making a decision. For example, in case of mammography, a CNN is trying to answer what features of a mammogram are most predictive of the presence of a cancer within.

RADIOMICS AND RADIOGENOMICS: SHIFT TOWARD PERSONALIZED PATIENT CARE

Images contain information far beyond what meets the eye. While radiologists can interpret some of these features with the naked eye (such as margins, heterogeneity, density etc), pixel-by-pixel analysis of these images can yield significant amounts of hidden information. Studies have shown that these may be successfully correlated to outcomes such as patient survival and genomic mutations [5]. More specifically, it was shown by Choudhery *et al* [6] that in addition to differentiating among the molecular subtypes of breast cancer, texture features including entropy were significantly different among HER2 positive tumors showing complete response to chemotherapy. Other parameters such as standard deviation of signal intensity were found predictive in triple negative cancers. A similar study by Chen *et al* [7] in patients with non-small cell lung cancer treated with chemoradiotherapy showed that the 'radiomics signature' could predict failure of therapy. Therefore, using non-invasive imaging, it is thus possible to predict the mutations, response to specific drugs, best site of biopsy. Thus potentially, the therapy of the patient can be guided by markers mined from non-invasive imaging, making precision medicine a true possibility.

DISCUSSION

Common Applications of AI in Oncology

Most applications of AI in oncology are currently in the field of radiology and pathology, given the abundant digital data available in these fields. These tasks may be classified into specific categories (Figure 4). For readers wishing to foray into the field, an explanation of each kind of task, as well as data requirements and some examples of applications of these tasks are given below.

Classification

A classification task is one in which the AI algorithm classifies each image as belonging to one of several target categories. These categories are given at the level of image or patient. For instance, whether a particular mammogram has cancer or not.

Data requirement: Training the network requires input images (mammograms in the above example), and an image level ground truth label (presence or absence of cancer in the above case). These are relatively easy to obtain if reports are available in a digital format, since automated extraction of diagnosis from free-text reports may be performed. Usually thousands of such images are required for training. Large public datasets of labelled natural images exist, such as "ImageNet" with over 14 million images [8], and several classification networks trained on these databases also exist, such as Alexnet, Inception, ResNet *etc.* These networks trained on these large public databases can be adapted to the medical domain, a process called "Transfer Learning".

Classification tasks can be evaluated by calculating the Area Under Receiver Operating Curve (AUROC) and by drawing a confusion matrix from which accuracy of classification can be calculated.

Applications: Some examples of classification tasks include breast density categorization on mammograms [9], detection of stroke on head CT in order to prioritize their reading [10], prioritising chest radiographs based on presence of pneumothorax in them [11].

Advantages: The most advantageous use of classification networks is for triage. These can be used to classify images that need urgent attention, or those that need a re-look by a reporting radiologist, pathologist or ophthalmologist. This helps to reduce workload and effectively divert resources where required.

Disadvantages: When a classification task is performed by an algorithm, it simply classifies an image into a certain category, say 'benign' or 'malignant' for a mammogram, or 'COVID' or 'Non COVID' for a chest radiograph. It does not indicate which part of the image it used for classification, or indeed, if multiple lesions were present which lesion it classified. This translates to reduced 'explainability' of such a model, where the results cannot be understood logically.

Detection

A detection task is one in which the network would predict the presence as well as location of a lesion on an image. Unlike a classification task, which is performed at image or patient level, the detection task is performed at lesion level. For example if the network draws a box around a cancer on a mammogram, the task is a detection task.

Data Requirement: Training requires images as input, the ground truth needs to be provided as a box (called a bounding box) around each lesion, with their labels mentioned. This would typically have to be done prospectively, as this is not performed in the routine work-flow of most departments. In the above example, each mammogram would have to be annotated with bounding boxes by an expert radiologist (usually by multiple radiologists to avoid missing/ misclassifying lesions), and each box would have to be assigned a label (as benign/malignant or with a BIRADS score, depending on what output is expected). Several publicly available datasets such as the COCO dataset [12] exist for natural images, with several networks trained on these datasets for object detection (such as RCNN, faster-RCNN, YOLO etc). Detection tasks are evaluated by calculating the intersection over union (IOU) between a predicted box, and a ground truth box; that is by calculating how close a predicted box is in comparison to the ground truth box. All boxes over a certain cut-off are considered a correct prediction. A Free-Response Operating Curve (FROC) is drawn and sensitivity of the network at specific false positivity rates can be computed and compared.

Applications: The most prominent applications in oncology are detection of nodules on chest radiographs [13] and CT scans of the lungs [14-16], and detection of masses and calcifications on mammography [17].

Segmentation and Quantification

A lesion segmentation task essentially involves classifying each pixel in the image as belonging to a certain category. So unlike a classification task (image or patient level) or detection task (lesion level), a segmentation task is performed at pixel level. For instance, classification of each pixel of a CT image of the liver as background liver or a lesion would result in demarcating the exact margins of a lesion. The volume of these pixels may then be calculated, and give the volume of the right lobe, left lobe of the liver separately.

Data Requirement: Here, exact hand annotations of the lesion in question by the expert is required. This involves drawing an exact boundary demarcating the exact lesion in each section of the scan. Since this is routinely performed for radiotherapy planning, such data may be leveraged for building relevant datasets. Datasets like the COCO dataset exist with pixel level annotations for natural images.

These algorithms are evaluated with segmentation accuracy, IOU with the ground truth annotations (described in the previous section) or Dice scores [18].

Advantages: There is tremendous advantage to the use of AI for segmentation, particularly quantification, in terms of increasing throughput and reducing the man-hours required for these tasks. In some cases such as quantification of extent of emphysema, which is particularly tedious for human operators, ready acceptance of AI may be found.

Applications: Automated liver volume calculations (liver volumetry) is an important application which can significantly reduce the time of the radiologist spent in the process [19,20]. Segmentation of cerebral vessels to perform flow calculations [21], segmentation of ischemic myocardial tissue [22] are other such applications.

Image generation

Image generation refers to the network “drawing” an image, based on images it has seen. For instance, if a network is trained with low dose CT and corresponding high resolution CT images, it may learn to faithfully draw the high-resolution CT image, given the low dose CT. The most successful neural network to perform this task is called a Generative Adversarial Network (GAN), first described by Ian Goodfellow [23]. This involves training 2 CNNs- a Generator, which draws the image, and a discriminator, which determines whether a given image is real or generated. The 2 CNNs are trained simultaneously, with each trying to get better than the other.

Data requirement: This kind of network is usually trained in an “unsupervised” manner, that is, no ground truth is required. Therefore no expert time is required in annotating these images. Only curated datasets of a particular kind of images are required.

This kind of network is difficult to evaluate, since no objective measure is typically defined. Evaluation by human eyes is generally considered the best

Applications: GAN has found use in several interesting and evolving applications. This includes CT and MRI reconstruction techniques to improve spatial resolution while reducing the radiation dose or time of acquisition respectively. GANs can also be trained to correct or remove artifacts from images [24]. An interesting application of GAN has been in generating images of a different modality, given an image of a certain modality. An example is generation of a PET image, given a CT image [25], generation of MRI brain image from CT brain [26], or a T2 weighted image from T1 weighted image [27].

Advantages: An interesting application of GAN has been used for simulation training for diagnostic imaging [28,29]. Students may be trained to recognise a wide variety of pathology using the synthetic images generated from these networks. This may be particularly important in certain scenarios such as say, detecting masses in dense breasts.

Disadvantages: These networks seem to possess a supra-human capability. The generated images cannot be verified; for authenticity of texture or indeed even representation and thus may lead to an inherent mistrust of 'synthetic' images.

Natural Language Processing (NLP)

NLP refers to understanding of natural human language. While processing structured information is relatively easy, most data in the real world is locked up in the form of sentences in natural language. For example, understanding what is written in radiology reports would require processing of free-text, this task is called NLP.

Data requirement: Large publically available datasets such as the "Google blogger corpus" (text) and "Spoken Wikipedia corpuses" (spoken language) are available, over which networks can be trained to understand natural language. However large medical corpuses with reports pertaining to specific tasks are needed for tackling specific medical problems. With more robust Electronic medical records (EMR), integrated Hospital and Radiology Information Systems (HIS and RIS) as well as recorded medical transcripts, this field is likely to grow rapidly.

Applications: The applications of NLP range from extraction of clinical information from reports and EMRs to train deep neural networks, to designing chatbots for conversing with patients.

Predictive Modelling, Radiomics and Radiogenomics

Predictive modelling has been at the core of medical practice for decades. While initial attempts were centered at developing scoring systems, or metrics that could be calculated from a few lab parameters, predictive modelling can be much more complex today because of the number of variables that machine learning systems can analyse.

A simple example of such a model is the “cholesterol ratio” (Total cholesterol/ HDL) which is used to estimate the risk of cardiac disease. As our models are capable of processing many variables, in fact capable of processing whole images, predictive models can be much more nuanced. Radiomics and Radiogenomics are in fact an extension of the same, built to predict survival, response to therapy or future risk of cancer, with more complex feature extraction and analysis from radiology images

Data requirement: Building such models requires longitudinal data. Simple machine learning models would require lesser data in comparison to deep learning models. The amount of data required essentially depends upon which level machine learning is used at. For instance, if lesion segmentation is performed manually, feature extraction is performed with routine textural features, and feature selection is performed by means of traditional tools such as simple clustering or principle component analysis (PCA), then machine learning model would only use these selected features to make the desired prediction, and the amount of data required is relatively small. However if deep learning is used end-to-end, the data requirement is much higher.

Predictive models are also assessed through AUROC and confusion matrices from which accuracy of prediction can be calculated.

Radiomics involve 4 steps: a) segmentation b) extraction of features c) selection of features and d) model building for prediction **(Figure 5**). Segmentation involves drawing a margin around a lesion. This may be performed by an expert manually, or automatically. Features of the lesion are then defined. These may be semantic, that is defined by an expert, such as tissue heterogeneity, spiculated margins etc, or quantitative features (such as mean, median, histogram analysis, filter-extracted features). This may yield several 100 features, of which overlapping features should be removed before analysis. Subsequently a few selected features may then be fed into a machine learning model along with the outcomes that are to be predicted. Machine learning or Deep learning may also be applied at the initial stages, for segmentation and feature extraction itself, rather than at the last step.

Applications: Predictive models are extremely useful in oncology. Studies have shown that features extracted from images can be used to predict the response to various kinds of therapies. Morshid *et al* and Abajian *et al* showed good accuracy at predicting response to transarterial chemoembolisation (TACE) [30,31]. Studies have also correlated the imaging features extracted with genomic information, for example, several studies have shown that imaging features can accurately predict EGFR mutation status in patients with lung cancer [32-35], Segal *et al* showed that 28 CT texture features features could decode 78% of the genes expressed in hepatocellular carcinoma [36]. More recent work also shows that deep learning models can predict future risk of development of cancer. Eriksson *et al* studied a model that identified women at high likelihood of developing breast cancer within 2 years based on the present mammogram [37]. All these pave the way towards more personalised management of patients with cancer.

7. Genomic data analysis

The next generation of personalised medicine is undoubtedly 'genomic medicine', wherein not just targeted therapy, but also diagnostic procedures are tailored as per the genetic make-up of an individual.

In addition, there is a growing effort towards population based studies for pooling of large scale genomic data and understanding the relationship between genomics, clinical phenotype, metabolism and such domains. The challenge with these techniques are the huge amounts of data obtained from a single cycle, and the computational requirements in its processing and analysis. Thus, both ML and DL are ideally suited to deal with each step of the process starting from genome sequencing to data processing and interpretation.

For instance, a deep learning model that combined both histological and genomic data in patients with brain tumors to predict the overall survival, was able to show non-inferiority compared to human experts [38].

Data requirement: The essential in this field is not the data, but rather the ability to process the data. Since the human genome contains approximately 3 billion base pairs, and thousands of genes, the data becomes extremely high dimensional. Convolutional neural networks and recurrent neural networks (RNNs) have been proven to be the best approach to evaluate multiple DNA fragments in parallel, similar to the approach used in Next Generation Sequencing (NGS) [39]. RNN models have also been used to perform microRNA and target prediction from gene expression data [40].

Applications: In addition to the applications detailed above, AI has also found use in variant identification, particularly Google's 'Deep Variant' which has shown superior performance to existing methods despite not being trained on genomic data [41]. Other studies have also used machine learning to identify disease biomarkers and predict drug response [42,43].

ENABLING PATIENT-CENTRIC ONCOLOGY CARE

Much of medical care today is moving away from patients, with focus shifting towards interpreting digital data in the form of blood reports, imaging data, pathology reports, genomic information *etc.* While the sheer amount of data has rendered face to face patient care less important, as synthesizing this information takes significant time and effort.

ML and DL have however ushered in a new era with endless possibilities. For instance, in a field like radiologist, where AI is likely to have maximum impact, the onco-radiology reporting room of the future is likely to be dramatically different from where we are currently. AI, by reducing the amount of time spent in preparing a report, may pre-prepare images and sample reports, allowing a radiologist to spend time with the patient, examining the clinical files and providing the report immediately after the examination (unlike in current practice where a radiologist sees the images, never meets the patient and gives them a report about 24 h later). This report can potentially be transcribed into several reports simultaneously - for instance a patient friendly report, in easy to understand non-medical language, a physician report with important sections and lesions marked on the image, and a traditional descriptive radiology report. In fact the radiology report is likely to have much more information than currently considered possible, including the possibility of a particular mutation, possibility of response to a particular therapy, even reconstructed images translated to different modalities which may help determine the most important site of biopsy *etc.*

While Amara's law for new technology may well apply (which says that any new technology is overestimated early on, and underestimated later [44]), the potential of AI and the vistas that it open up cannot be ignored. As the technology evolves, many of the changes it brings about will enable a leap towards the era of personalised medicine. (**Figure 6**)

CONCLUSION

AI thus holds great potential. The most significant advantage of AI rests in the fact that since it is data-driven, it holds the potential to derive inferences from very large databases, in a short span of time. It brings with it the possibility to standardize clinical care, reduce interpretation times, improve accuracy of diagnosis, and may help enable patient centricity in cancer care.

Like any new technology, however, AI must be used with care and only after thorough clinical tests. The most significant disadvantage derieves from the fact that it is a "black-box", with little explainability . Little is known about the reasons behind the decisions taken by neural networks, making it imperative for the decisions to be seen and approved by human experts.

In summary, there is tremendous scope of artificial intelligence in cancer care, particularly in the image related tasks. With the development of neural networks capable of performing complex tasks, the era of personalised medicine seems a reality with AI. Thus, judicious use must be encouraged to maximise the long term benefits that outlive the initial enthusiasm of discovery.

75772_Auto_Edited.docx

ORIGINALITY REPORT

1 %

SIMILARITY INDEX

PRIMARY SOURCES

1	Ronak Bahuva, Joe Aoun, Sachin S. Goel. "Management of Acute Coronary Syndrome in the COVID Era", Methodist DeBakey Cardiovascular Journal Internet	12 words — < 1%
2	www.ijcseonline.org Internet	12 words — < 1%

EXCLUDE QUOTES	ON	EXCLUDE SOURCES	OFF
EXCLUDE BIBLIOGRAPHY	ON	EXCLUDE MATCHES	< 12 WORDS