Advances in Deep Learning for Cancer Diagnosis

Dilip Kumar Mishra, Sudhansu Sekhar Sahoo, Prasanta Kumar Sahoo, Prangya Paramita Padhi

Department of Electronics and Communication Engineering, NM Institute of Engineering and Technology,Bhubaneswar , Odisha
Department of Electronics and Communication Engineering, Raajdhani Engineering College,Bhubaneswar,Odisha
Department of Electronics and Communication Engineering,Aryan Institute of Engineering and Technology Bhubnaeswar , Odisha
Department of Electronics and Communication Engineering,Capital Engineering College,Bhubaneswar,Odisha

ABSTRACT: *Deep learning refers to a set of computer models that have recently been used to make unprecedented progress in the way computers extract information from images. These algorithms have been applied to tasks in numerous medical specialties, most extensively radiology and pathology, and in some cases have attained performance comparable to human experts. Furthermore, it is possible that deep learning could be used to extract data from medical images that would not be apparent by human analysis and could be used to inform on molecular status, prognosis, or treatment sensitivity. In this review, we outline the current developments and state-of-the-art in applying deep learning for cancer diagnosis, and discuss the challenges in adapting the technology for widespread clinical deployment.*

I. INTRODUCTION TO DEEP LEARNING

Deep learning (DL; see Glossary) consists of a set of machine learning algorithms, also known as deep neural networks (DNNs), that have achieved unprecedented success over the past decade in processing forms of natural data, such as images, text, and speech [1]. Machine learning, broadly speaking, applies statistical methods to training data to automatically adjust the parameters of a model, rather than a programmer needing to set them manually. Histori-cally, machine learning algorithms such as random forests and support vector machines have performed well with structured forms of data, but have struggled with data that does not have a consistent organization. Somewhat paradoxically, even years after defeating human chess grandmasters, it was still impossible for computers to perform image recognition tasks that would be trivial for a child. Slow progress was being made in this area, but in 2012, Krizhevsky et al. used a form of DNN called a convolutional neural network (CNN) to make a major jump in the accuracy of general image recognition [2], and since then, CNNs have become the de facto approach for most computer vision tasks (Box 1 and Figure 1). DL algorithms have further achieved success as critical elements in systems that play games that were thought to exemplify human intuition or instinct, such as the strategy game Go [3] and poker [4].

II. HIGHLIGHTS

Several factors, including advances in computational techniques and algo-rithms, the availability of graphical pro-cessing units, and the assembly of large training datasets, have led to the establishment of DL as the domi-nant method for computer vision tasks.

Competitions in image processing have focused effort on particular tasks and have been useful in revealing which approaches are the most successful.

DL algorithms have attained expert level performance in the detection of breast cancer metastases in lymph nodes, and demonstrated superior accuracy compared with previous fea-ture-engineered methods of histology analysis.

DL can facilitate large-scale morphol-ogy-based research, such as in the recent mapping and analysis of tumor infiltrating lymphocyte patterns in thou-sands of specimens from the Cancer Genome Atlas digital slide repository.

The theoretical underpinnings of DNNs have existed for decades, but several synergistic developments have led to a recent popularization [5]. Advances in the mathematical methods used to train DNNs, such as improvements in back-propagation, have addressed many of the issues that historically made them challenging to optimize, particularly as they become larger [6–9]. These models require large amounts of training data, and the proliferation of online databases over the past two decades has provided exponential increases in the amount of image and text data available. The widespread availability and affordability of graphical processing units (GPUs), largely driven by the video game industry, has provided the computational power needed to train DNNs in a reasonable time.

Box 1. Technical Overview of CNN Training

Traditional neural networks are composed of fully connected layers stacked from the input to the eventual output layer. CNNs are a form of neural network with three types of layers convolutional, pooling, and fully connected. The fully connected layers in a typical neural network are not optimized to realize local patterns that depend on the proximity of features, an important capability for image analysis. Convolutional layers overcome thischallenge by imitating the behavior of the human eye and sampling local spots of information thatoverlap to some extent. Computationally,this task is executed by splitting the input image into overlapping tiles that are defined by a manually selected filter size and stride. Each tile is then transformed into a single numerical value through multiplication with a kernel. The size of the resulting feature map can be further reduced bycombining adjacent tiles with a max-pooling layer,which keeps the maximumvalue from a set of adjacent tiles, thereby retaining maximal information from the area. An activation function applied to the kernel output introduces nonlinearity and ensures that values are comparable across tiles. This procedure of sequential convolution and pooling can be iterated to generate increasingly compact feature maps of the input data, the last of which serves as input for a fully connected set of layers. Each node (neuron) in these layers is connected to all the nodes in the preceding layer, and transforms their output by a set of weights, with the addition of a layer-specific constant value, called a bias term, to ensure that the output from the node is non-zero. The final layer is typically a softmax function that converts the activations of the preceding layer into a range of probabilities across the set of output classes. As with other neural networks, a CNN is trained end-to-end, from the input image to the output classification, using back-propagation. A preselected cost function calculates the error in the output, a measure of how far the predictions are from the ground truth. The optimization function, such as stochastic gradient descent, dictates how this cost propagates through the weights of each layer. Using back-propagation iteratively, the feature maps gradually shift to select features from the input that are increasingly informative for the classification task. Once the calculated cost becomes stable over multiple iterations of the training set, training can be stopped, and new samples predicted with the learnt weights. development of user-friendly, open source programming libraries like Keras and Tensorflow has significantly lowered the barrier to entry for non-computer scientists to engage in DL research [10] (Table 1).

Over the past several years, research into the medical applications for DL has accelerated, with cancer being the most common disease investigated and images the dominant data type [11]. The applications of DL for cancer diagnosis can be broadly divided into two uses that we label automated analysis and knowledge discovery. Automated analysis refers to the use of models for routine clinical diagnostic tasks, in which expert level performance has been reached in several medical fields [12–14], while knowledge discovery aims to uncover new patterns in data that may inform on diagnosis, prognosis, treatment response, or genomic status (Table 2). In this review, we summarize DL applications in cancer research pertaining to radiology and pathology, the image-based specialties that are involved in virtually every cancer diagnosis, and consider the future impact of artificial intelligence (AI) technology on medical practice.

DL in Radiology

The field of radiology has long been at the forefront of incorporating computers into clinical practice, beginning with their use for administration and billing in the 1960s [15]. Computed tomography (CT) and magnetic resonance imaging (MRI), which were both invented in the early 1970s and proliferated in the clinic through the 1980s, acquire images digitally; however, limitations in computer storage required hard copies of these scans to initially be developed on radiographic film [16]. The later development of Picture Archiving and Communication Systems (PACS) in the 1990s enabled the transition to a fully digital workflow, and today most radiographic imaging in Canada and the USA is obtained, viewed, and stored digitally. In the early 2000s, the US Congress approved the use of computer-aided diagnostics for screening mammography under Medicare coverage, as well as the replacement of transcrip-tionists by text recognition systems. Numerous subsequent studies into the clinical benefit of mammography screening have produced conflicting results as to their clinical benefit [17,18]. A sensitive system is generally desirable for screening-based tasks; however, a high rate of false-positives can distract the radiologist and potentially even lead to increased biopsies.

Glossary

Artificial intelligence (AI): use of computers to model some or all aspects of human intelligence. Includes DL and other machine learning methods, as well as previous knowledge-based approaches that attempted to hard code inference rules. Convolutional neural network (CNN): form of DL that is particularly well suited to image analysis. CNNs use alternating convolutional and pooling operations to extract spatially invariant features from input data, while limiting the number of parameters in the network.

Deep learning [DL; also known as deep neural networks (DNNs)]: form of machine learning that uses complex multilayered architectures to extract progressive degrees of abstraction from input data. End-to-end system: machine learning system that maps input data (after preprocessing) directly to predictions, without the use of a separate feature extraction step. Graphical processing units (GPUs): form of integrated circuit that has been designed specifically to efficiently alter memory for the display of computer graphics. Their highly parallelized structure is also efficient at the large-scale matrix operations that are used in neural networks.

Hand-engineered features (also referred to as hand-crafted features): features used for prediction that have been manually selected or inferred by the designer of the model.

Machine learning: application of statistical methods to adjust the parameters of a model based on training data, rather than being explicitly programmed. Preprocessing: transformations applied to an image prior to using it as input for a CNN, such as normalization, standardization, resizing, cropping, rotation, or color adjustment.

Radiomics: field of research that aims to extract minable data from radiographic imaging. Random forest: machine learning algorithm that uses a large number of individually weak decision trees to generate robust predictions.

While these initial forays into computer-aided diagnosis have not had widespread clinical uptake, it should be noted that they used technology that preceded the rise of DL, and recent head-to-head comparisons have demonstrated superior performance of DL over other sys-tems [19,20]. Several recent studies trained on large datasets have demonstrated comparable performance of DL systems to that of experts in common diagnostic tasks across a range of modalities, including chest X rays [21], head CT [22], spine MRI [23], mammography [20], and limb trauma X rays [24]. With the increasing evidence that CT chest screening can reduce lung cancer mortality, the automated detection and evaluation of lung nodules has generated considerable interest, including two large international challenge competitions [25][i].

Organ or lesion segmentation (the automated delineation between tissues and tissue struc-tures) is often a necessary initial step that supports both further analysis and some forms of treatment, making it a key piece of automated systems. There has been extensive work done in this area across a range of organs and pathology types [26]. Within this research there has been a particular focus on segmentation tasks in neuroimaging, including numerous challenge competitions involving brain tumors, non-neoplastic lesions, and normal brain structures [27]. Arterys, a San Francisco based startup, recently received FDA clearance for a suite of DL-based oncology image analysis products[ii], the first such approval. The software currently focuses on lung and liver analysis, with approval to ultimately expand to all solid tumors, and is able to segment lesions, track them across time, and assist with common radiological scoring systems.

In addition to uses that directly impact diagnoses, DL has other applications that can improve the radiology workflow, including image quality enhancement, alignment of multiple images, content-based retrieval, report generation and semantic error flagging, and database mining for research [26,28]. Outside of cancer specific diagnosis, a DL-based system to triage CT head scans for radiologist review based on the presence or absence of critical findings, has demonstrated utility in a simulated clinical environment by decreasing the time taken for radiologists to review the more urgent images [29].

Knowledge discovery in radiology largely falls under the field of radiomics (or radiogenomics); the field that aims to extract minable data from imaging. In the past, approaches based on the use of a small number of hand-engineered features have identified imaging based correlates of molec-ular subtype and prognosis in numerous cancer types, including breast cancer [30], glioblastoma [31], renal cell cancer [32], and head and neck squamous cell carcinoma [33]. Similar work incorporating DL methods has demonstrated promising results in areas such as the prediction of isocitrate dehydrogenase (IDH), 1p19q, and O^6-methylguanine-DNA-methyltransferase (MGMT) status in gliomas [34–36], malignant potential in gastrointestinal stromal tumors [37], and of breast cancer molecular subtype [38] based on imaging. Ongoing work is needed to further elucidate the roles of hand-crafted features versus end-to-end DL systems in this field, as they will likely be complementary approaches.

III. DIGITAL PATHOLOGY AND MORPHOLOGICAL ANALYSIS

In contrast to the digitization of clinical radiology practice, pathology continues to predomi-nantly use glass slides and light microscopes, even in the most advanced hospitals. Recent advances in scanner technology and storage capacities have made it feasible for laboratories to implement fully digital systems, and several laboratories have published case studies describ-ing their transition [39–41]. In Canada, whole slide imaging (WSI) has been used to facilitate intraoperative frozen sections for rapid diagnosis and consultations for over a decade, and was approved for primary diagnosis in 2013 [42,43]. However, in the USA progress has been in part

Segmentation: delineation of different normal and/or abnormal structures or regions. Examples of segmentation tasks include outlining different organs on radiographic images and mapping regions of invasive cancer on histology. Support vector machine: machine learning algorithm that uses a set of hyperplanes to distinguish between classes of data with the widest possible margin.

Whole slide imaging (WSI): use of advanced scanning technology to scan entire glass histology slides at a sufficiently high resolution for pathologic analysis (typically 200–400 X magnification).

Max-pooling across

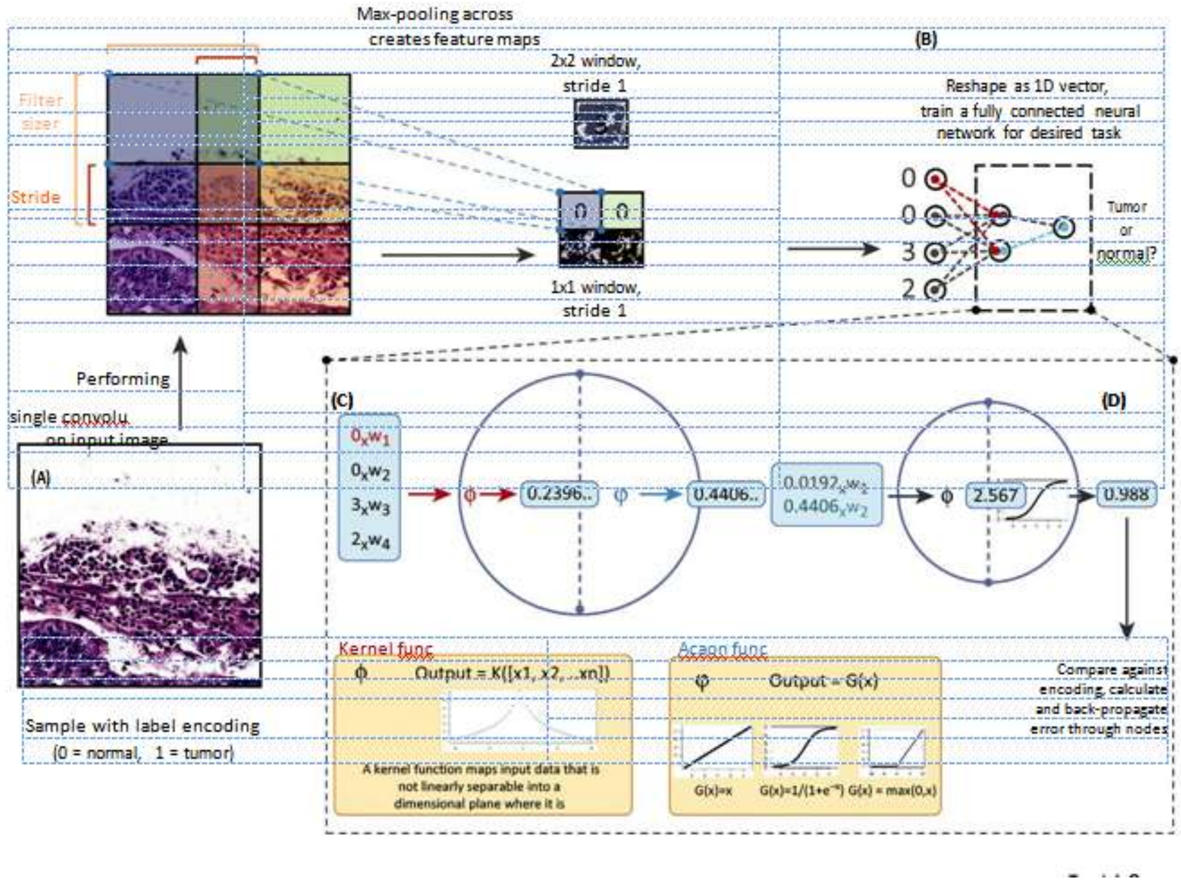

Figure 1. Visualization of Convolutional Neural Network Layers and Functions. (A) An input image undergoes several rounds of convolution and pooling operations to extract progressively higher order features. (B) Following these, the feature maps are reshaped as a 1D vector and fed into a fully connected layer, which outputs the final prediction. (C) At each layer, the weights kernel is applied to the image, and the resulting value is run through an activation function (typically a rectified linear unit, or ReLU for short). (D) In the final layer a softmax function is applied, which generates probabilities for each output class. These are then compared with the ground-truth label to determine the error of the prediction. hampered by the regulatory environment, in which WSI systems have been considered Class III medical devices[iii] – the highest risk devices, requiring rigorous premarket approval [44,45].

An important moment occurred in 2017, with FDA approval of the Phillips Intellisite as the first WSI system to be used for primary diagnosis. This approval was based on a study involving 16 pathologists across four sites that demonstrated equivalent error rates between manual diagnoses made on digital and glass slides[iv]. Since then, one American laboratory has already described their transition to WSI for primary diagnosis[v]. A fully digitized laboratory has numer-ous potential benefits for patient safety, workflow efficiency, and quality improvement [46], but challenges to implementation include the need for internal WSI system validation, high upfront cost and additional personal requirements, and pathologist acceptance [45]. Our hope is that, as the

Internet has provided huge amounts of general images to train DL models, the expanding access to WSI will be a similar catalyst for the growth of pathology specific applications. That said, it is crucial to recognize that slides that are digitized for clinical purposes contain confidential patient information as metadata bundled with the file, as well as in scanned labels within the image. This information must be appropriately removed prior to use for research, particularly in slides that are to be released as publicly available datasets or used by private companies.

Table 1. Landmark DL Papers and Other Important Resources

Year	Importance	Refs
1986	First publication of the back-propagation method for adjusting the parameters in a neural network based on the gradient of the error.	[89]
1990	Earliest use of a CNN trained by back-propagation, in this case to recognize handwritten digits.	[90]
2012	Landmark paper in which a CNN nearly halved the previous best error rate for image recognition. This was largely what precipitated the current rise of DL.	
2015, 2016	Papers presenting the Inception and Resnet CNN architectures, which both achieved state of the art results in image recognition. Many of the papers cited in this review used modified versions of the models, including those attaining expert level performance	
2015	Tensorflow and Keras are two of the most widely used software libraries[ix,x] for training neural networks.	
2016	Review of DL written by three pioneers in the field.	[1]
2016	Comprehensive textbook of current DL methods and research.	[5]

As noted, radiographic images are predominantly acquired digitally, allowing for the transition to a fully digital workflow with minimal loss of information. Tissue sections, in contrast, potentially contain more information than might be obtained in a digital image, particularly given the ability to vary depth of focus and apply high magnification to selected regions. The potential loss of fine detail, combined with concerns that digital slides take longer to review than glass ones, has reduced the attractiveness of WSI to pathologists [45]. Scanning capacity has only recently reached an appropriate scale to manage clinical slide volume. For example, the Vancouver General Hospital (VGH)[vi], a typical large tertiary center, produces about 400 000 stained sections per year. A modern slide scanner requiring 2 min per slide can theoretically scan about 200 000 slides per year, although practically, this throughput may not be reached due to ongoing challenges in identifying and focusing on tissue regions, and interference from artifacts such as tissue folding and air bubbles. In addition to the capital and service costs of scanners, enterprise quality data storage would cost around US$100 000 per year to store all

Table 2. Key Papers Applying DL to Cancer Diagnosis

Year	Medical field	Task	Refs
2017	Dermatology	Classification of benign versus malignant skin lesions	[14]
2017	Pathology	Detection of breast cancer metastases in lymph nodes	[12]

2018	Pathology	Prediction of glioblastoma survival from histology	[63]
2018	Pathology	Analysis of tumor infiltrating lymphocyte patterns in 13 cancer types and correlation with molecular markers and survival	[63]
2017	Radiology	Detection of pulmonary nodules on chest CT	[25]
2017	Radiology	Detection of 14 pathologies, including lung nodules and masses, on chest X ray.	[21]

the slide output of VGH. These costs and challenges suggest that clinical scale WSI will only become compelling to most hospitals when it exhibits significant benefits.

Improvements in slide imagers have motivated the development of a range of tools designed to make diagnosis and grading less subjective by quantifying image features known to correlate with disease state [47]. In tumor pathology, for instance, where nuclear morphology and cellular architecture are often strong determinants of disease severity, algorithms can be designed to detect dysplasia or invasive tumors by first segmenting nuclei from background, quantifying a number of nuclear features, such as size, shape, and spacing, and comparing these features with those typical of normal cells [48]. This approach has generated good results across many tissue types, but has been particularly successful in cytology [49] and hematology [50], where the segmentation of single cells on a homogeneous background is less challenging [47]. Developments in image processing and statistical methods have enabled greater sophistica-tion in the design of these algorithms, and a state-of-the-art algorithm might use thousands of features to derive its predictions [51].

Despite their continued improvements, feature-based algorithms often suffer from two limi-tations. The first is a lack of consistency in the performance of the same algorithms with runs on sections prepared with different staining protocols or scanned under different conditions. Algorithms that depend on accurate segmentation can be sensitive to changes in color and brightness and can yield inconsistent results on samples from different centers, even when stain normalization is used [52]. Other preanalytic variables that can influence algorithm performance include tissue quality, fixation, slice thickness, and any artifacts (glue, air bubbles, etc.). The second issue is that these algorithms rely on a prespecified set of features to classify the tissue. Because they can only classify tissue as well as the features that distinguish between them, there is a ceiling to their performance, even when a large amount of data is available to refine the algorithm [12].

DL in Pathology

DL provides a significantly different approach to histopathology image analysis than feature-based methods. As end-to-end systems, DL systems dispense with the initial feature extraction step. Instead, after basic preprocessing, images are fed directly into the model, which by virtue of a large parameter space incorporates its own automated feature extraction into the earlier layers of the network. This approach requires modification when large, high-resolution images are used. Whole digital slides, unlike other common medical image types, can be >1 GB each, which is too large to be processed by the model in their entirety. Instead, the typical approach is to crop the slide into numerous small image patches; process these as essentially independent of each other; and then aggregate the patch-level predictions to make an overall slide-level prediction or a heatmap of regions of interest (Figure 2). Early work in 2014 using this approach showed promising results in the identification of invasive ductal breast carcinoma [53].

In a seminal paper in 2017, Ehteshami Bejnordi et al. published the results of an international competition in the identification of metastatic breast cancer deposits in lymph nodes [12]. This was an ideal task for initial medical applications of DL, as it is well defined, repetitive, and high volume, yet potentially error prone for humans. Twenty-three different teams submitted predictions on a test set of 129 WSIs of lymph nodes, which were compared against two benchmarks set by human experts. A panel of pathologists with a soft time constraint of 2 h was used to approximate real world performance, while a single pathologist without time constraints, who spent over 30 h evaluating the slides, provided an estimate for the upper limit of human performance. The top algorithms had similar results as the pathologist without time constraints, and generally exceeded those of the panel pathologists. These results provide the strongest evidence to date that DL models have the potential to reach expert pathologist level performance; albeit on one narrow task. An important caveat is that these algorithms were only trained on metastatic cancer and likely would not have detected other pathology that can be present in lymph nodes, such as lymphomas or reactive conditions. While this study did not include external validation, an algorithm developed using this dataset has since been validated on slides from an independent laboratory [54] and demonstrated clinical utility by improving pathologist accuracy and efficiency in detecting metastases on digital slides [55]. The follow-up challenge in 2017, extended this work to a more clinically realistic scenario involving the nodal staging of simulated patients comprised of sets of five slides [56].

Numerous other studies have applied DL to similar tasks in pathology. Other work in breast cancer has included the segmentation of tumor regions in breast resection slides [57], differentiation between several different types both of benign breast changes and cancer histotypes [58], and the identification of cancer based solely on alterations of the surrounding stroma [59]. Another high-volume, repetitive task that is well suited to automation is the evaluation of prostate biopsies and resection specimens, and preliminary work has demonstrated histological grading of tissue microarray specimens with interrater variability between the computer and two reference pathologists similar to that between the pathologists themselves [60]. In this study, visualization of the most salient features used by the model to make predictions confirmed that it was focusing on the epithelium, with a particular emphasis on the junctions between glands. More recently, a team from Google published their large-scale study on the scoring of prostate cancer on prostatectomy specimens [61]. DL can also be used to quantify important features in slides, with extensive work having been done in mitosis identification [62]; a particularly challenging task in WSI, given the lack of 3D information (the z axis).

In the area of knowledge discovery, the Cancer Genome Atlas (TCGA)[vii] digital slide repository has already proven to be a rich resource for combining histology with clinical and molecular data, leading to several high-quality publications. Mobadersany et al. developed what they have termed a survival CNN in order to predict glioma outcomes [63]. Based on histology alone, their model was able to differentiate outcomes within molecular subtypes of glioma, while it obtained improved prognostic accuracy by combining histology with common genomic markers. Heat map visualizations indicated that higher risk was predicted in regions with conventionally malignant histological features, as well as in regions with previously unrecognized features, such as adjacent regions of edema and sparsely infiltrated brain, illustrating the potential of DL to identify useful features that could be added to routine histological evaluation by pathologists. It should be noted that this study was only evaluated on TCGA slides and it remains to be seen whether the prognostic significant of the algorithm applies to external data sets.

Saltz et al. used a CNN-based computational staining methodology to map the patterns of tumor-infiltrating lymphocytes in over 5000 slides across 13 cancer types and correlate it with molecular subtypes and survival [64]. They used an iterative process to develop their network, in which a limited number of slides were annotated and used to train the initial model, and the predictions of this model were then corrected by pathologists and fed back in as additional training examples. This was repeated until a satisfactory performance was reached; at which point the model could be deployed on the full dataset. In addition to its insights on the immune response within tumors, this studyprovides a blueprint for theuse ofautomated image processing tofacilitate morphology-based research on a scale that would not be feasible if pathologists had to annotate every slide.

In prostate cancer, DL was used to automate the identification of the most abnormal regions on slides (analogous to what is done manually for tissue microarray construction) in order to predict speckle-type POZ protein (SPOP) status [65]. This study trained the model on frozen slides from the TCGA archive but then tested institutional paraffin embedded tissue, demonstrating consistency of the algorithm despite varying slide quality. Furthermore, the authors used an innovative strategy to address the challenge of dataset imbalance with rare mutations, by forming an ensemble of multiple models trained on subsets of the data with matched numbers of positive and negative slides. Similarly, in lung cancer, DL has been used to predict the status of several driver mutations in adenocarcinoma [66], as well as overall outcomes based on morphological features [67]. DL has also been used to model clinical behavior from genomic profiling [68], which could be combined with image analysis to further refine these predictions [69].

IV. FUTURE PROSPECTS AND CHALLENGES

The rise of AI has unquestionably been a disruptive force in a number of industries and is poised to cause even more disruption. This potential has inevitably and understandably led to clashing viewpoints as to its role and incorporation in society in the future. Unlike most historical technological advancements, which have predominantly affected manual work, AI is expected to have a significant impact on so-called knowledge workers. In a survey that asked several hundred machine learning experts about the effect of AI on a range of jobs, the median prediction was that AI will outperform humans in performing surgery by the year 2053 (with a range of 2030–2100), just later than the predicted time AI will be able to write a bestselling novel, but earlier than that predicted for performing mathematical research [70]. There is robust debate among pathologists as to the projected future role of human specialists and the potential for AI to exceed human diagnostic capabilities [71,72]. However, in considering these issues, it is important to remember the inherent imprecision of technological prognosti-cation and the role of perspectives and biases in influencing individual opinions [73].

The research discussed in this review has certainly been promising, and demonstrated convincingly that in some tasks AI can match the performance of human medical experts. Beyond the ongoing work in further optimizing DL algorithms, there are significant barriers to adapting this technology into widespread medical use and to truly approximate the cognitive processes of a human physician. Most DL uses have been highly task specific, while humans are able to make associations that can improve performance across multiple related tasks. Despite the limitations of feature-engineered approaches, there will likely be benefits in combining semantic knowledge with visual analysis, particularly in distinguishing between rare diagnoses with limited examples available. Furthermore, conclusions in radiology and pathol-ogy are often not based just on a single scan or specimen, but also on correlation with previous ones and other medical history [72].

DL is in general data hungry – significantly more so than earlier feature-engineered approaches that are less prone to overfitting – and the acquisition of sufficient training data is an ongoing challenge in nearly all domains. While unsupervised and semisupervised learning approaches exist, for most medical tasks, data sets require manual annotation or at least curation [74]. Depending on the complexity of the task this may be appropriate for trained research personnel or require the full input from medical experts. As the early layers in DNNs almost invariably learn very general image features, networks that have been pretrained on large general image sets can be fine-tuned on medical data, which can decrease the amount of data required and overall training time [75,76]. Furthermore, novel methods are being developed to facilitate slide annotation, such as incorporation directly into the clinical workflow by tracking pathologist movements as they read slides [77], or by combining expert and crowd-sourced annotations [78].

In comparison to the feature-engineered approaches that have been discussed, DL has been criticized for being a 'black box', in which it is not entirely clear how the model generates outputs from a given input. While this argument certainly has some merit, methods to visualize the activation functions of a network and the types of images that activate a given neuron have helped to elucidate the inner workings of these algorithms, and this remains an area of active research [79,80]. An analogy can be drawn between the interpretability issue in DL and FDA-approved drugs with unknown mechanisms of action [81], as well as our incomplete understanding of the human cognitive diagnostic process [82]. Regardless, given the inability of current DL algorithms to explain their diagnostic process, several issues would need to be addressed prior to their implementation in clinical practice, including the degree of physician supervision that is required and determining who is ultimately liable for machine error. In this regard, cues can potentially be taken from the similarly high-risk field of autonomous driving, where five levels of system autonomy have been defined, ranging from basic driver assistance to full automation without human backup [83].

Despite the hype and high expectations surrounding DL in medicine, it is crucial that medical regulators and practitioners proceed with caution and insist that new algorithms are rigorously validated in realistic environments prior to use for patient care [84]. One particular challenge of regulating AI algorithms is that they are not static products, and can continue to change and improve even once deployed, as new training data becomes available. At this point, the FDA regulates DL-based algorithms as medical devices[viii], and several have been approved for radiology in the past 2 years, but none for pathology image analysis. The FDA has signaled plans to streamline its process for approval of AI algorithms, but it is still unclear what precise regulatory framework will enable the rapid advances in this field while maintaining patient safety [85,86]. For the foreseeable future it is likely that AI will remain in a diagnostic support role, in which it can help detect pathologies, automate routine tasks, and improve workflow, but a human will retain responsibility for all final decisions and reports. In Figure 3, we illustrate a hypothetical use of AI in the management of a patient with a brain tumor, from the initial radiographic imaging to the pathology report from the tumor resection. However, given the current state of the field, specific details as to the implementation of this technology remain largely speculative.

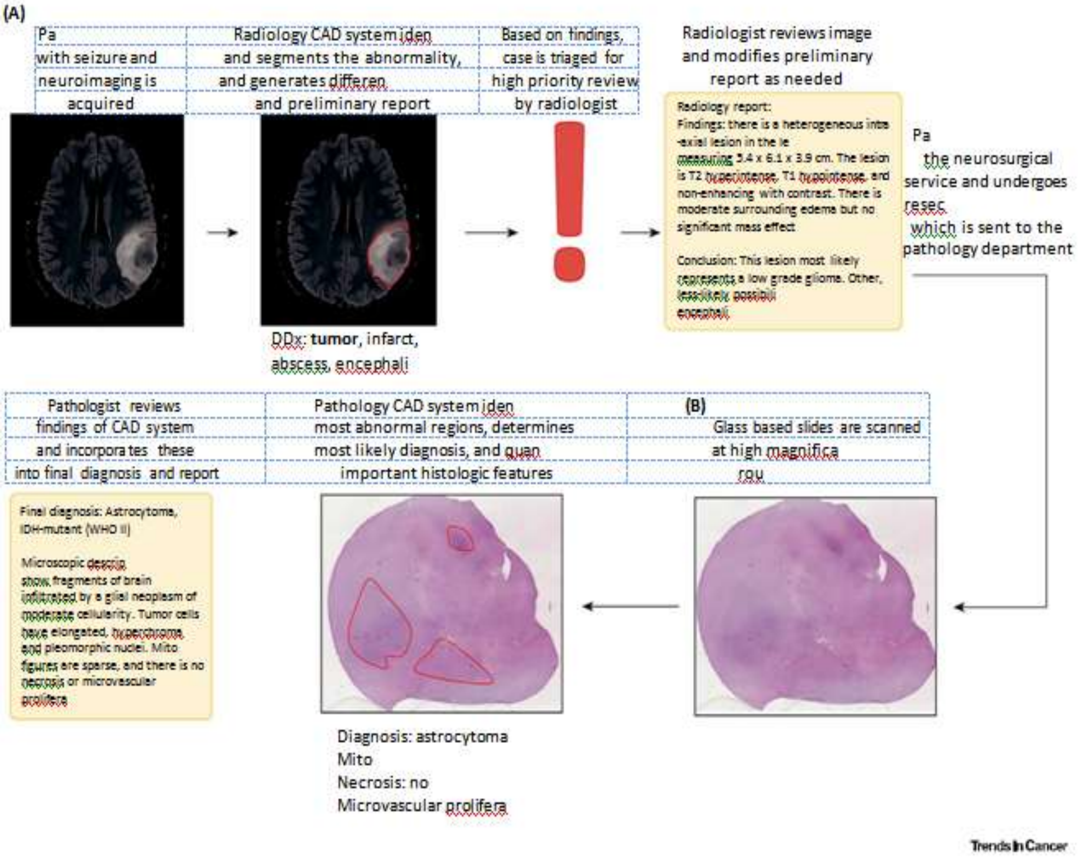
(A)
Pa
with seizure and
neuroimaging is
acquired
Radiology CAD system iden
and segments the abnormality,
and generates differen
and preliminary report
Based on findings,
case is triaged for
high priority review
by radiologist
Radiologist reviews image
and modifies preliminary
report as needed
Radiology report:
Findings: there is a heterogeneous intra
-axial lesion in the le
measuring 5.4 x 6.1 x 3.9 cm. The lesion
is T2 hyperintense, T1 hypointense, and
non-enhancing with contrast. There is
moderate surrounding edema but no
significant mass effect
Conclusion: This lesion most likely
represents a low grade glioma. Other,
less likely possibili
encephali
Pa
the neurosurgical
service and undergoes
resec
which is sent to the
pathology department
DDx: tumor, infarct,
abscess, encephali
Pathologist reviews
findings of CAD system
and incorpora tes these
into final diagnosis and report
Pathology CAD system iden
most abnormal regions, determines
most likely diagnosis, and quan
important histologic features
(B)
Glass based slides are scanned
at high magnifica
rou
Final diagnosis: Astrocytoma,
IDH-mutant (WHO II)
Microscopic descrip
show fragments of brain
infiltrated by a glial neoplasm of
moderate cellularity. Tumor cells
have elongated, hyperchrome
and pleomorphic nuclei. Mito
figures are sparse, and there is no
necrosis or microvascular
prolifera
Diagnosis: astrocytoma
Mito
Necrosis: no
Microvascular prolifera
Trends in Cancer

Figure 3. Proposed Method of Incorporating Artificial Intelligence into Diagnostic Medicine Workflow.

(A) A patient's initial magnetic resonance imaging scan is analyzed by a computer-aided diagnostic system, which generates a preliminary report and flags it for high priority review by a radiologist. (B) The resected tumor specimen is received in the laboratory, and the glass slides are digitized as part of the workflow. These are then analyzed by the computer system, and its findings integrated with those of the pathologist to generate a final report. Abbreviations: CAD, computer-aided diagnosis; DDx, differential diagnosis.

Regardless of the eventual impact of DL specifically, the practice of diagnostic medicine will continue to change as new technologies are introduced. If DL algorithms are able to generate widespread clinician acceptance, the cost of the computational infrastructure needed to deploy DL algorithms (as opposed to train them) is likely minimal in the context of overall healthcare spending [28]. Should AI ultimately be able to automate a good portion of image analysis, the job of a radiologist or pathologist may shift to increasingly emphasize other tasks, such as correlation with the medical records, formulating reports, liaising with clinicians, departmental quality control, and participating in multidisciplinary conferences. This technology will also necessitate a shift in the training of diagnostic physicians to better understand the computa-tional techniques involved, with some suggesting the creation of an entirely new specialty or even a merger of pathology and radiology [87,88].

Concluding Remarks

In summary, DL is an exciting development in the ongoing pursuit of computer-aided medical diagnostics. Research over the last several years indicates its potential to attain human expert level performance, but the technology appears to remain distant from widespread clinical deployment (see Outstanding Questions). AI will likely change the practice of diagnostic medicine, and we are optimistic that it will ultimately lead to improved patient safety and quality of medical care.

Outstanding Questions
How can the process of annotating training data be better integrated into the clinical workflow?
Which clinical tasks are appropriate for DL?
How much data is needed for any par-ticular DL task?
How can multiple research groups be coordinated to assemble high-quality datasets?
Can a DL system be designed that will have broad task capability?
How can DL and hand-engineered fea-tures best be combined?
Will the molecular and clinical predic-tions generated by DL be clinically useful?
How can we better understand the mechanics through which DL systems generate predictions?
How will DL diagnostic systems be regulated and what will be the required level of human oversight?
Who will be responsible for medical errors made by this technology?
What will be the impact of AI on physi-cian employment?
How will society react towards the implementation of AI systems in medicine?

REFERENCES

[1]. LeCun, Y. et al. (2015) Deep learning. Nature 521, 436–444

[2]. Krizhevsky, A. et al. (2012) ImageNet classification with deep convolutional neural networks. Proceedings of the 25th Interna-tional Conference on Neural Information Processing Systems 1, pp. 1097–1105

[3]. Silver, D. et al. (2016) Mastering the game of Go with deep neural networks and tree search. Nature 529, 484–489

[4]. Moravcik, M. et al. (2017) DeepStack: expert-level artificial intelli-gence in heads-up no-limit poker. Science 356, 508–513

[5]. Goodfellow, I. et al. (2016) Deep Learning, MIT Press

[6]. Glorot, X. et al. (2011) Deep sparse rectifier neural networks. Proceedings of the 14th International Conference on Artificial Intelligence and Statistics, pp. 315–323

[7]. Ioffe, S. and Szegedy, C. (2015) Batch normalization: accelerating deep network training by reducing internal covariate shift.

[8]. Proceedings of the 32nd International Conference on Interna-tional Conference on Machine Learning 37

[9]. He, K. et al. (2016) Deep residual learning for image recognition. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778

[10]. Szegedy, C. et al. (2015) Going deeper with convolutions. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1–9

[11]. Erickson, B.J. et al. (2017) Toolkits and libraries for deep learning. J. Digit. Imaging 30, 400–405

[12]. Jiang, F. et al. (2017) Artificial intelligence in healthcare: past, present and future. Stroke Vasc. Neurol. 2, 230–243

[13]. Ehteshami Bejnordi, B. et al. (2017) Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. JAMA 318, 2199–2210

[14]. Gulshan, V. et al. (2016) Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA 316, 2402–2410

[15]. Esteva, A. et al. (2017) Dermatologist-level classification of skin cancer with deep neural networks. Nature 542, 115–118

[16]. Mayo, R.C. and Leung, J. (2018) Artificial intelligence and deep learning – radiology's next frontier? Clin. Imaging 49, 87–88

[17]. Rubin, G.D. (2014) Computed tomography: revolutionizing the practice of medicine for 40 years. Radiology 273, S45–74

[18]. Fenton, J.J. et al. (2007) Influence of computer-aided detection on performance of screening mammography. N. Engl. J. Med. 356, 1399–1409

[19]. Lehman, C.D. et al. (2015) Diagnostic accuracy of digital screen-ing mammography with and without computer-aided detection. JAMA Intern. Med. 175, 1828–1837

[20]. Wang, X. et al. (2017) Searching for prostate cancer by fully automated magnetic resonance imaging classification: deep learning versus non-deep learning. Sci. Rep. 7, 15415

[21]. Kooi, T. et al. (2017) Large scale deep learning for computer aided detection of mammographic lesions. Med. Image Anal. 35, 303–312

[22]. Rajpurkar, P. et al. (2017) CheXNet: radiologist-level pneumonia detection on chest X-rays with deep learning. arXiv:1711.05225v1

[23]. Merkowa, J. et al. (2017) DeepRadiologyNet: radiologist level pathology detection in CT head images. arXiv:1711.09313v3

[24]. Jamaludin, A. et al. (2017) ISSLS Prize in Bioengineering Science 2017: automation of reading of radiological features from mag-netic resonance images (MRIs) of the lumbar spine without human intervention is comparable with an expert radiologist. Eur. Spine J. 26, 1374–1383

[25]. Olczak, J. et al. (2017) Artificial intelligence for analyzing ortho-pedic trauma radiographs. Acta Orthop. 88, 581–586

[26]. Setio, A.A.A. et al. (2017) Validation, comparison, and combina-tion of algorithms for automatic detection of pulmonary nodules in computed tomography images: the LUNA16 challenge. Med. Image Anal. 42, 1–13

[27]. Litjens, G. et al. (2017) A survey on deep learning in medical image analysis. Med. Image Anal. 42, 60–88

[28]. Akkus, Z. et al. (2017) Deep learning for brain MRI segmenta-tion: state of the art and future directions. J. Digit. Imaging 30, 449–459

[29]. Liew, C. (2018) The future of radiology augmented with artificial intelligence: a strategy for success. Eur. J. Radiol. 102, 152–156

[30]. Titano, J.J. et al. (2018) Automated deep-neural-network surveil-lance of cranial images for acute neurologic events. Nat. Med. 24, 1337–1341

[31]. Mazurowski, M.A. et al. (2014) Radiogenomic analysis of breast cancer: luminal B molecular subtype is associated with enhance-ment dynamics at MR imaging. Radiology 273, 365–372

[32]. Gutman, D.A. et al. (2013) MR imaging predictors of molecular profile and survival: multi-institutional study of the TCGA glioblas-toma data set. Radiology 267, 560–569

[33]. Karlo, C.A. et al. (2014) Radiogenomics of clear cell renal cell carcinoma: associations between CT imaging features and muta-tions. Radiology 270, 464–471

[34]. Leger, S. et al. (2017) A comparative study of machine learning methods for time-to-event survival data for radiomics risk model-ling. Sci. Rep. 7, 13206

[35]. Akkus, Z. et al. (2017) Predicting deletion of chromosomal arms 1p/19q in low-grade gliomas from MR images using machine intelligence. J. Digit. Imaging 30, 469–476

[36]. Korfiatis, P. et al. (2017) Residual deep convolutional neural network predicts MGMT methylation status. J. Digit. Imaging 30, 622–628

[37]. Li, Z. et al. (2017) Deep learning based radiomics (DLR) and its usage in noninvasive IDH1 prediction for low grade glioma. Sci. Rep. 7, 5467

[38]. Ning, Z. et al. (2018) Pattern classification for gastrointestinal stromal tumors by integration of radiomics and deep convolu-tional features. EE J. Biomed. Health Inform. Published online May 29, 2018. http://dx.doi.org/10.1109/JBHI.2018.2841992

[39]. Zhu, Z. et al. (2017) Deep Learning for identifying radiogenomic associations in breast cancer. arXiv:1711.11097

[40]. Cheng, C.L. et al. (2016) Enabling digital pathology in the diag-nostic setting: navigating through the implementation journey in an academic medical centre. J. Clin. Pathol. 69, 784–792

[41]. Stathonikos, N. et al. (2013) Going fully digital: perspective of a Dutch academic pathology lab. J. Pathol. Inform. 4, 15

[42]. Thorstenson, S. et al. (2014) Implementation of large-scale rou-tine diagnostics using whole slide imaging in Sweden: digital pathology experiences 2006-2013. J. Pathol. Inform. 5, 14

[43]. Tetu, B. et al. (2014) The Eastern Quebec Telepathology Network: a three-year experience of clinical diagnostic services. Diagn. Pathol. 9, S1

[44]. Tetu, B. and Evans, A. (2014) Canadian licensure for the use of digital pathology for routine diagnoses: one more step toward a new era of pathology practice without borders. Arch. Pathol. Lab. Med. 138, 302–304

[45]. Parwani, A.V. et al. (2014) Regulatory barriers surrounding the use of whole slide imaging in the United States of America. J. Pathol. Inform. 5, 38

[46]. Griffin, J. and Treanor, D. (2017) Digital pathology in clinical use: where are we now and what is holding us back? Histopathology 70, 134–145

[47]. Williams, B.J. et al. (2017) Future-proofing pathology: the case for clinical adoption of digital pathology. J. Clin. Pathol. 70, 1010–1018

[48]. Gurcan, M.N. et al. (2009) Histopathological image analysis: a review. IEEE Rev. Biomed. Eng. 2, 147–171

[49]. Diamond, J. et al. (2004) The use of morphological characteristics and texture analysis in the identification of tissue composition in prostatic neoplasia. Hum. Pathol. 35, 1121–1131

[50]. Guillaud, M. et al. (2006) DNA ploidy compared with human papilloma virus testing (Hybrid Capture II) and conventional cer-vical cytology as a primary screening test for cervical high-grade lesions and cancer in 1555 patients with biopsy confirmation. Cancer 107, 309–318

[51]. Briggs, C. et al. (2009) Can automated blood film analysis replace the manual differential? An evaluation of the CellaVision DM96 automated image analysis system. Int. J. Lab. Hematol. 31, 48–60

[52]. Yu, K.H. et al. (2016) Predicting non-small cell lung cancer prog-nosis by fully automated microscopic pathology image features. Nat. Commun. 7, 12474

[53]. Yoshida, H. et al. (2018) Automated histological classification of whole-slide images of gastric biopsy specimens. Gastric Cancer 21, 249–257

[54]. Cruz-Roa, A. et al. (2014) Automatic detection of invasive ductal carcinoma in whole slide images with convolutional neural net-works. SPIE Med. Imaging 9041, 15

[55]. Liu, Y. et al. (2018) Artificial intelligence-based breast cancer nodal metastasis detection. Arch. Pathol. Lab. Med. Published online October 8, 2018. http://dx.doi.org/10.5858/arpa.2018-0147-OA

[56]. Steiner, D.F. et al. (2018) Impact of deep learning assistance on the histopathologic review of lymph nodes for metastatic breast cancer. Am. J. Surg. Pathol. 42, 1636–1646

[57]. Bandi, P. et al. (2019) From detection of individual metas-tases to classification of lymph node status at the patient level: the CAMELYON 17 challenge. IEEE Trans. Med. Imag-ing 38, 550–560

[58]. Cruz-Roa, A. et al. (2017) Accurate and reproducible invasive breast cancer detection in whole-slide images: a deep learning approach for quantifying tumor extent. Sci. Rep. 7, 46450

[59]. Habibzadeh Motlagh, N. et al. (2018) Breast cancer histopatho-logical image classification: a deep learning approach. bioRxiv 242818

[60]. Ehteshami Bejnordi, B. et al. (2018) Using deep convolu-tional neural networks to identify and classify tumor-asso-ciated stroma in diagnostic breast biopsies. Mod. Pathol. 31, 1502–1512

[61]. Arvaniti, E. et al. (2018) Automated Gleason grading of prostate cancer tissue microarrays via deep learning. Sci. Rep. 8, 12054

[62]. Nagpal, K. et al. (2018) Development and validation of a deep learning algorithm for improving Gleason scoring of prostate cancer. arXiv:1811.06497

[63]. Veta, M. et al. (2015) Assessment of algorithms for mitosis detec-tion in breast cancer histopathology images. Med. Image Anal. 20, 237–248

[64]. Mobadersany, P. et al. (2018) Predicting cancer outcomes from histology and genomics using convolutional networks. Proc. Natl. Acad. Sci. U. S. A. 115, E2970–E2979

[65]. Saltz, J. et al. (2018) Spatial organization and molecular correla-tion of tumor-infiltrating lymphocytes using deep learning on pathology images. Cell Rep. 23, 181–193 e7

[66]. Schaumberg, A.J. et al. (2017) H&E-stained whole slide image deep learning predicts SPOP mutation state in prostate cancer. bioRxiv 064279

[67]. Coudray, N. et al. (2018) Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. Nat. Med. 24, 1559–1567

[68]. Wang, S. et al. (2018) Comprehensive analysis of lung cancer pathology images to discover tumor shape and boundary fea-tures that predict survival outcome. Sci. Rep. 8, 10393

[69]. Chaudhary, K. et al. (2018) Deep learning-based multi-omics integration robustly predicts survival in liver cancer. Clin. Cancer Res. 24, 1248–1259

[70]. Savage, R.S. and Yuan, Y. (2016) Predicting chemoinsensitivity in breast cancer with ‘omics/digital pathology data fusion. R. Soc. Open Sci. 3, 140501

[71]. Grace, K. et al. (2018) When will AI exceed human performance? Evidence from AI experts. arXiv:1705.08807v3

[72]. Granter, S.R. et al. (2017) AlphaGo, deep learning, and the future of the human microscopist. Arch. Pathol. Lab. Med. 141, 619–621

[73]. Sharma, G. and Carter, A. (2017) Artificial intelligence and the pathol-ogist: future frenemies? Arch. Pathol. Lab. Med. 141, 622–623

[74]. Granter, S.R. (2016) Reports of the death of the microscope have been greatly exaggerated. Arch. Pathol. Lab. Med. 140, 744–745

[75]. Hosny, A. et al. (2018) Artificial intelligence in radiology. Nat. Rev. Cancer 18, 500–510

[76]. Yosinski, J. et al., How transferable are features in deep neural networks? Proceedings of the 27th International Conference on Neural Information Processing Systems, Montreal, Canada, 2014, pp. 3320–3328

[77]. Kajbakhsh, N. et al. (2016) Convolutional neural networks for medical image analysis- full training or fine tuning? IEEE Trans. Med. Imaging 35, 1299–1312

[78]. Schaumberg, A.J. et al. (2017) DeepScope: nonintrusive whole slide saliency annotation and prediction from pathologists at the microscope. Comput. Intell. Methods Bioinform. Biostat. (2016) 10477, 42–58

[79]. Albarqouni, S. et al. (2016) AggNet: deep learning from crowds for mitosis detection in breast cancer histology images. IEEE Trans. Med. Imaging 35, 1313–1321

[80]. Zintgraf, L.M. et al. (2017) Visualizing deep neural network deci-sions: prediction difference analysis. arXiv:1702.04595

[81]. Zeiler, M.D. and Fergus, R. (2013) Visualizing and understanding convolutional networks. arXiv:1311.2901

[82]. Gregori-Puigjane, E. et al. (2012) Identifying mechanism-of-action targets for drugs and probes. Proc. Natl. Acad. Sci. U. S. A. 109, 11178–11183

[83]. Bruno, M.A. et al. (2015) Understanding and confronting our mistakes: the epidemiology of error in radiology and strategies for error reduction. Radiographics 35, 1668–1676

[84]. Holzinger, A. et al. (2017) Towards the augmented pathologist: challenges of explainable-AI in digital pathology. arXiv:1712.06657v1

[85]. Topol, E.J. (2019) High-performance medicine: the convergence of human and artificial intelligence. Nat. Med. 25, 44–56

[86]. Allen, B. (2019) The role of the FDA in ensuring the safety and efficacy of artificial intelligence software and devices. J. Am. Coll. Radiol. Published online October 30, 2018. http://dx.doi.org/10.1016/j.jacr.2018.09.007

[87]. Pesapane, F. et al. (2018) Artificial intelligence as a medical device in radiology: ethical and regulatory issues in Europe and the United States. Insights Imaging 9, 745–753

[88]. Lundstrom, C.F. et al. (2017) Integrated diagnostics: the compu-tational revolution catalyzing cross-disciplinary practices in radi-ology, pathology, and genomics. Radiology 285, 12–15

[89]. Jha, S. and Topol, E.J. (2016) Adapting to artificial intelligence: radiologists and pathologists as information specialists. JAMA 316, 2353–2354

[90]. Rumelhart, D.E. et al. (1986) Learning representations by back-propagating errors. Nature 323, 533–536

[91]. LeCun, Y. et al. (1990) Handwritten digit recognition with a back-propagation network. Proceedings of Advances in Neural Infor-mation Processing Systems, pp. 396–404