Deep learning for image analysis: Personalizing medicine closer to the point of care

Quin Xie, Kevin Faust, Randy Van Ommeren, Adeel Sheikh, Ugljesa Djuric & Phedias Diamandis

Published online: 10 Jan 2019.

Submit your article to this journal

View Crossmark data

Taylor & Francis
Taylor & Francis Group

Check for updates

REVIEW ARTICLE

Deep learning for image analysis: Personalizing medicine closer to the point of care

Quin Xie[a,b]*, Kevin Faust[b,c]*, Randy Van Ommeren[a,b]*, Adeel Sheikh[b], Ugljesa Djuric[b] and Phedias Diamandis[a,b,d]

[a]Department of Laboratory Medicine and Pathobiology, University of Toronto, Toronto, Canada; [b]MacFeeters-Hamilton Brain Tumour Centre, Princess Margaret Cancer Centre, Toronto, Canada; [c]Department of Computer Science, University of Toronto, Toronto, Canada; [d]Laboratory Medicine Program, University Health Network, Toronto, Canada

ABSTRACT

The precision-based revolution in medicine continues to demand stratification of patients into smaller and more personalized subgroups. While genomic technologies have largely led this movement, diagnostic results can take days to weeks to generate. Management at, or closer to, the point of care still heavily relies on the subjective qualitative interpretation of clinical and diagnostic imaging findings. New and emerging technological advances in artificial intelligence (AI) now appear poised to help bring objectivity and precision to these traditionally qualitative analytic tools. In particular, one specific form of AI, known as deep learning, is achieving expert-level disease classifications in many areas of diagnostic medicine dependent on visual and image-based findings. Here, we briefly review concepts of deep learning, and more specifically recent developments in convolutional neural networks (CNNs), to highlight their transformative potential in personalized medicine and, in particular, diagnostic histopathology. Understanding the opportunities and challenges of these quantitative machine-based decision support tools is critical to their widespread introduction into routine diagnostics.

Introduction

Providing the right treatment to the right patient at the right time represents the essential ingredients of personalized medicine [1,2]. While genomic medicine is proving to be a formidable tool toward resolving the former two components, in many medical centers, molecular results take days to weeks to report, limiting their utility for guiding clinical decision making in acute and primary care settings. As many personalized management plans could benefit from early initiation of therapy, there is a clinical need and opportunity to deliver precision-based results at the bedside, or at least closer to the point of care (Figure 1).

Most diagnostic information in the acute setting, however, is largely available in the form of visual cues or medical images rather than discrete and objective numerical values. For instance, in primary care, physicians rely on qualitative physical findings to evaluate and assess the severity of a wide variety of apparent changes to health (e.g. dermatological lesions, jaundice, and breast masses). Similarly, other patient concerns

(e.g. chest pain) are assessed using visual interpretation of image data in the form of electrocardiograms (ECGs) at the point of care. Even when examined by specialists, bedside sub-classification of patients' symptoms into actionable entities still heavily rely on subjective and qualitative visual interpretations rather than objective quantitative values. These "pattern-recognition" tasks occur at multiple levels within diagnostic and treatment workflows, including initial presentation, physical examination, medical imaging, initial pathologic interpretation, and even during surgical management. Moreover, inter-observer variability in tests ordered and their interpretation, challenges diagnostic precision. In the end, even after multiple diagnostic procedures, serial confirmatory testing is often required. These iterative steps are costly, delay initiation of optimal care, and lead to significant emotional distress to patients and their families. Personalization of medicine at, or closer to, the point of care thus requires a balance between early detection of emergent diseases (sensitivity) while reducing the health, economic, and patient
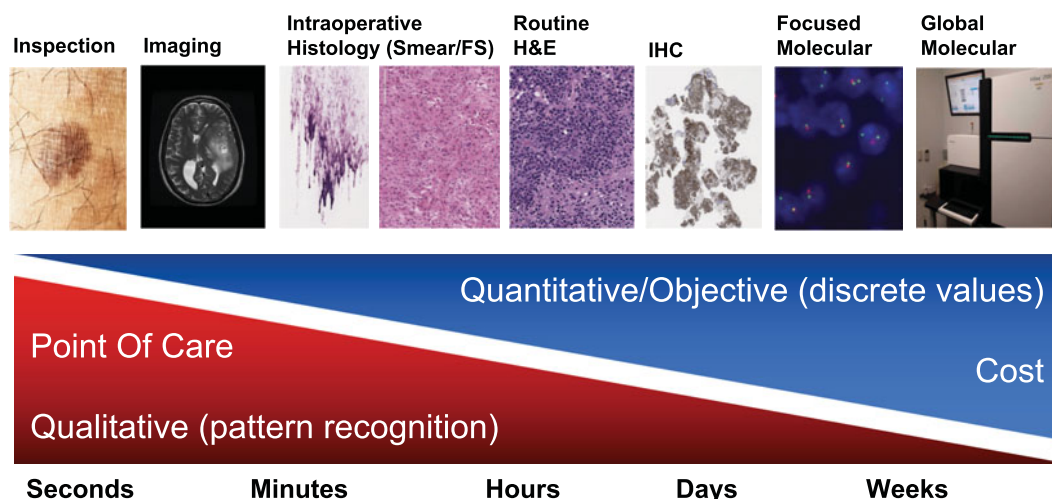
Figure 1. Pattern recognition tasks in diagnostic medicine. Unlike the non-acute setting where medical information is usually generated as discrete, objective and quantitative values, most clinical data accessible early and close to the point of care, is typically visual. This information often requires highly-trained personnel for interpretation and even then, classifications are usually broad, qualitative and prone to inter-observer variability. FS: frozen section; H & E: hematoxylin and eosin; IHC: immunohistochemistry.

psychological burden of lengthy and comprehensive invasive testing (specificity) [3]. These important parameters associated with the value of a diagnostic procedure are often compromised with qualitative tests. Development of more objective analytical tools for the acute setting thus represents an area of tremendous opportunity for precision medicine [4].

Toward this, recent advances in artificial intelligence (AI) now allow computers to carry out image-based pattern-recognition classification tasks in a quantitative manner. One form of AI in particular, known as deep learning, uses an algorithm structure known as a convolutional neural network (CNN) to excel at preserving and analyzing spatially dependent information. In recent years, this technology has helped personalize many aspects of everyday life in real-time. This includes object and facial recognition in personal photo libraries, speech recognition for voice-based commands, and text string recognitions in e-mail messages for automatic calendar updates and spam prevention. These successes have sparked interest in potential applications of this form of AI for improving the objectivity and diagnostic yield of image-based tasks in medical diagnostics. Here we introduce some fundamental concepts of deep learning and describe selected recent triumphs in augmenting medical decision making in the more acute and sub-acute setting. Integration of these technologies into medical care may complement molecular medicine and allow deployment of the "right treatment" to the "right patient" in the timeliest and most cost-effective way possible.

Artificial intelligence

Understanding the position of deep learning in the broader category of AI and machine learning is paramount for appreciating its transformative power in medical pattern recognition tasks. Machine learning is a broad branch of computer science where large datasets are analyzed and used to build mathematical models that can make reasonable predictions for future cases. These are now ubiquitously used for generating multi-parametric classification signatures in the era of big data. For instance, recent abundance of DNA, RNA, and proteomic datasets [5–10] composed of quantitative multiple gene and protein values across large sample cohorts have long relied on machine learning technologies to find diagnostic and prognostic expression patterns (Figure 2(A)). Traditional machine learning approaches (e.g. Logistic Regression, Support Vector Machines, and Random Forests) have been critical for identification of multi-gene or protein signatures that indicate disease prognosis, which could not be analyzed manually by a human despite lengthy professional training. These tools have been very effective for interrogating large datasets to identify relationships between quantifiable data types. Traditional approaches to image recognition tasks involve the use of feature extractors to transform raw image data, such as pixel values, into numerical representations. Hundreds to thousands of such "engineered features" are then used to classify image patterns and correlate them with outcome or diagnostic labels. Unlike abundance levels of different molecules, however, visual
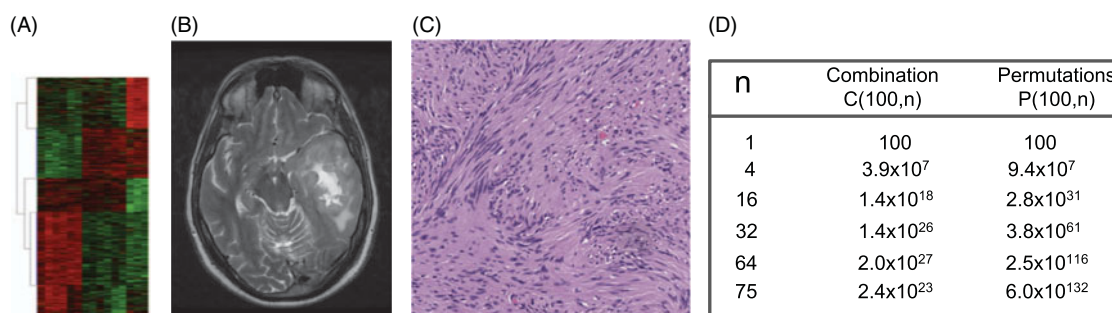
Figure 2. Computational analysis in medical data. Different data structures provide different orders of magnitude of computational complexity. (A) Traditional machine learning approaches are able to efficiently handle combinations of gene and protein expression value patterns where each data variable can be considered independent from one another. Permutational data, however, where data points are temporally or spatially (Panel B&C, MRI image and histology images) organized, pose progressively more challenging computational tasks. Panel D shows a table highlighting possible combination and permutations of a hypothetical task of arranging "n" data points chosen from a set of 100 to illustrate this. When the positional order of data matters (e.g. images), the number of possible arrangements quickly grows orders of magnitude higher than other data types.

information is dependent on the spatial coordinates of pixel values which are intimately tied to neighboring ones with significant implications on the meaning of the overall interpretation (Figure 2(B–D)). Moreover, variations in objects and localization in different positions within an image further complicates image-based classification tasks. These distinct differences in data structure, mathematically analogous to the differences between "combinations" and "permutations" of objects (data points), makes image analysis an exponentially more computationally challenging undertaking than traditional computational and pattern-recognition tasks in medicine and biology (Figure 2(D)).

Deep learning and convolutional neural networks

Previous advances in representation learning have promoted development of automated detection and extraction of specific features from images that can act as inputs for future segmentation and classification tasks. Evolution of these efforts has given rise to deep-learning approaches, which utilize learning architectures known as CNNs. Neural networks are constructed to simulate the layered connectivity of the human cortex, where layers of nodes simulate the connected neurons of the brain. These architectures permit highly complex levels of data representation, creating millions of parameters for effective image classification. Each layer of the neural net is composed of non-linear modules, transforming the representation at one level to a more abstract level in successive layers (Figures 3 and 4(B)). Layers at higher tiers amplify aspects that are crucial for discrimination (e.g. vertical and diagonal lines, circles, etc.) and diminish variations that may give false outputs (e.g. nonspecific variations in brightness). The

clear distinction of deep-learning from other machine learning processes involves the integration of multiple layers that are not programmed by human input (i.e. hand-crafted features); but instead are adopted by this model through a "data-driven" process. Each layer, and their interactions with one another, thus provides another layer of an expansive permutation of dynamic "features" by which learning can develop. Although a much more data demanding process, learning in this architecture theoretically continues to improve with the available data and has dominated other machine learning approaches for pattern recognition tasks (Figure 4(B,C)). Advancements in the parallel processing power of graphical processing units (GPUs) has now helped address the computational demands of these computer algorithms, making it practical to complete much larger numbers of computations with orders of magnitude of improvements in analytical speeds over traditional computer processing unit (CPU)-based computing.

Amidst all architectures applying deep learning for computer vision, CNNs have received the most attention due to their promising performance on visual recognition tasks. Instead of matching the whole image or object, as traditional machine learning methods do, CNNs match segments of images. CNNs combine multiple layers of representation learning with each layer gathering features ranging from the presence or absence of edges at the bottom layer to complex and abstract features of the overall pattern at the deeper layers (Figure 4(A)) [11]. In convolutional layers, training data is delivered from one layer to the next through filter banks and sets of shared weights are assigned to reduce the number of free parameters without loss of generalizability (Figure 4(A)). The pooling layers aggregate similar features by keeping the maximum or the average step size. The dropout layer applies regulator
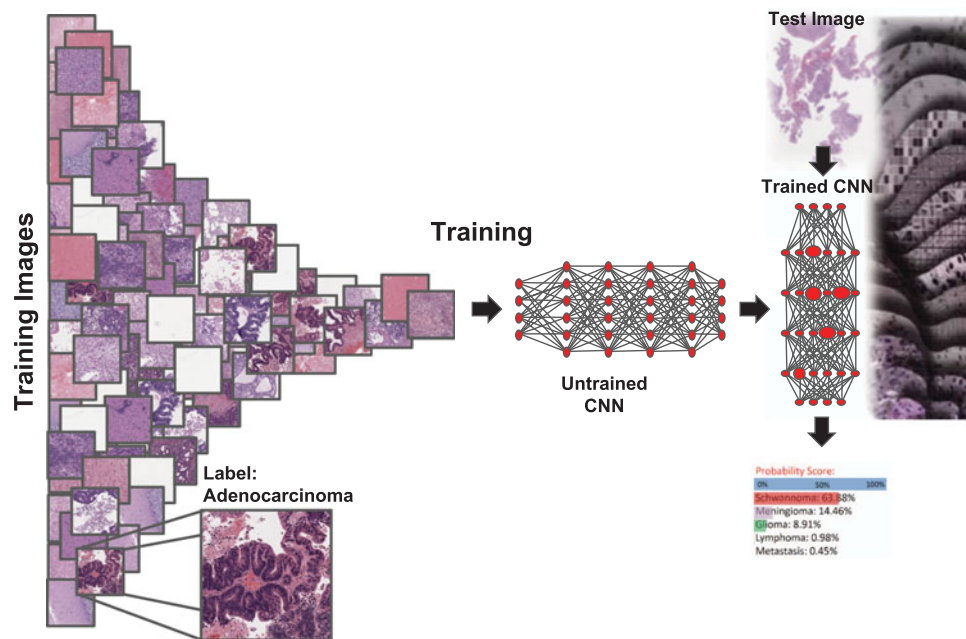
Figure 3. Precision in medical image analysis using deep neural networks. Advancements in computer vision now aim to bring objectivity to image analysis. Computational approaches, such as convolutional neural networks (CNNs), are now able to accept massive cohorts of annotated images and discover informative patterns in a data-driven process independent of hand-crafted features. These features are depicted here as different weightings (size of red nodes) at the different "nodes" founds in the CNN layers. Once training is complete, connections between the nodes of different layers work cooperatively to feature complex features used to classify future cases. Based on the sum of these features, test images can be quantitatively converted into probability distribution scores that provide diagnostic predictions.

functions, such as "normalization" and "rectification," to increase samples and to prevent overfitting. Softmax functions are usually applied in the output layer to prevent outliers from skewing the result. Driven by gradient descent, a value on the performance curve that attempts to minimize the error, this process facilitates analysis of objective visual-spatial pattern recognition and interpretation of images, providing an outlook not easily influenced by external factors. This unique approach has led to superior performance of CNNs in image-based classification tasks compared to other traditional machine learning approaches. Since 2011, no other machine learning approach has outperformed deep learning in the annual object recognition "ImageNet" challenge where computers are asked to classify objects spanning 1000 different classes [11–13] (Figure 4(C)). In these controlled environments, CNNs now achieve "super human" performance. Here we highlight how these very same concepts are beginning to be exploited for imaged-based classification tasks in medicine and particularly in pathology. These approaches are now beginning to bring objectivity and precision to relatively qualitative diagnostic decision making carried out closer to the point of care [14]. We use these examples, and our own experience, to discuss the potential utilities (and drawbacks) of

deep learning and how it stands to transform diagnostic medicine.

Deep learning in the primary and emergent care settings

Physical examination and diagnostic imaging provide valuable "actionable" information for healthcare practitioners. For example, in dermatology, inspection of skin lesions by a clinician permits the stratification of lesions into various categories (e.g. malignant, benign, suggestive of underlying genetic disorders like neurofibromatosis, etc.). However, the task remains subjective and imperfect for human observers, and given biases of human observers, prone to interobserver variability (Figure 1). Recently, a landmark study demonstrated that a robust skin lesion classification tool could be developed to take advantage of a pre-trained CNN further optimized with a smaller but specific database of approximately 130,000 clinical images relevant to 2032 different dermatological diseases. This tool can determine subtypes of skin lesions from photographic and dermoscopic images [3] with board-certified dermatologist-level accuracy on separating benign, malignant, and non-neoplastic skin lesions (∼72% vs. ∼66%, CNN vs. physician, respectively), and on separating nine

Figure 4. Image analysis using deep convolutional neural networks. (A) In deep learning, training images are first individually analyzed pixel by pixel to find simple elementary patterns, such as lines and circles. These features are then fed to deep "hidden" layers where these objects are integrated and used to create more complex features. Feature selection occurs through a process called back propagation where features of diagnostic importance are prioritized by changing "weightings" within the neural networks. By searching for these features in future images, computers can provide objective diagnostic interpretation of spatially dependent data. (B) Cartoon showing that, unlike traditional machine learning approaches that generally use dozens to hundreds of human "hand-crafted features" for pattern recognition, convolutional neural networks are able to develop orders of magnitude larger numbers of "data-driven" features that improve performance. This is a highly data intensive approach and usually requires large amounts of data before performance benefits can be achieved. As the layers of a neural network are increased, more complex features can be generated and extracted which leads to massive improvements in performance. This emerging pattern recognition tool now dominates computer vision. (C) This theoretical concept was supported in the ImageNet challenge where deep neural networks have dominated the competition since their first introduction in 2012. Since then, deep convolutional neural networks (green) have outperformed more shallow and traditional computer vision approaches (blue bars). Within each bar, the winning error rate and number of layers with the CNN are noted.

disease classes (~55% vs. ~54%, CNN vs. physician, respectively). When compared to the average performance of 21 dermatologists, the algorithm demonstrated higher sensitivity and specificity for recognizing keratinocyte carcinoma and melanoma. Given this high level of performance, such CNNs provide an extremely valuable tool for earlier detection and monitoring of skin lesions. Intriguingly, these results suggest that, in addition to controlled use in the clinical setting, with further optimization and validation, image-driven CNN could allow development and reliable deployment to people's personal devices, providing cost-effective and powerful tools for early screening and triaging of visible patient symptoms and complaints.

Analogous approaches have begun to yield results in other clinical specialties. Ophthalmic examination of the fundus is routinely performed by ophthalmologists to assess for presence of diabetic retinopathy. Recently, a pre-trained CNN fine-tuned on 128,175 retinal retinopathy images was able to categorize diabetic retinopathy grade at a similar level to ophthalmologists, achieving an area under the receiver operator curve (AUC) of over 99% on two-validation sets [15]. In neurology, deep learning approaches have begun to be developed for automatic detection of ictal activity on encephalographic (EEG) data for seizure patients. A 13-layer neural network trained on normal, pre-ictal, and ictal EEG data (100 EEG signals per class) was able to achieve 88.7% accuracy, with a specificity and sensitivity of 90.0 and 95.9%, respectively. Of note, this did not exceed performance of older machine learning approaches, which the authors attribute to the limited

training data set size of their study [16]. In cardiology, automated ECG analysis has been used in the clinical setting for some time, but often criticized for inaccuracies. Recently, a 34-layer CNN trained on 64,121 ECG records from 29,163 patients were able to outperform six independent cardiologists on a 336-case test set, in both sensitivity and precision, suggesting that currently used automatic ECG analysis may soon be supplanted by deep learning based approaches [17]. Finally, novel approaches are being developed to assist clinicians with less formalized, but critical, diagnostic interpretations made based on physicians' immediately available visual cues and experienced clinical judgment. For example, Face2Gene, a deep learning based technology, performs facial analysis on patients to identify dysmorphic features and predict various genetic disorders. This technology is currently utilized in genetics clinics around the world, and has been used for identification of common genetic disorders (e.g. Down Syndrome), and for prediction of genetic subtypes of less common disorders, such as glycosylphosphatidylinositiol biosynthesis defects (GPIBDs) [18,19]. Other applications, such as deep learning classifiers used in real-time during colonoscopy to assess colonic lesions (e.g. polyps), have been developed, and show considerable promise [20].

Medical imaging is another essential modality for diagnosis and treatment, often at point of care. X-ray is one of the first and most common imaging techniques used to triage nonspecific symptomatology in the emergency department. However, radiologists may reach significantly different conclusions or often, cannot reach a definitive diagnosis. Although double reading is an intuitive solution, diagnostic delays, as well as the associated increased labor and costs would benefit from an alternative computer-aided approach. To this end, the field of "radiomics" aims to apply traditional machine and deep learning models to radiology. Recently, a 121-layer CNN was developed that leveraged a training set of over 100,000 frontal-view X-ray images that spanned 14 diseases. During testing, the trained CNN exceeded the average performance of four practicing radiologists [21]. Another recent study [22] leveraged approximately 700 X-ray images to train computers to differentiate between benign and malignant lesions in mammography. Despite the relatively small number of images, the CNN model used by this group achieved an AUC of 0.822, outperforming previous machine learning classifiers for this problem (max AUC of 0.787). Although these results are not yet sufficient for true clinical implementation, a CNN-driven model for breast cancer detection with mammography is expected to improve with higher resolution images

and larger training sets. Additionally, computed tomography (CT) and MRI are other common and relatively noninvasive imaging modalities that provide diagnostic support to clinicians. Recently, deep learning of brain imaging procedures has been shown to be an effective approach for screening for neurodegenerative diseases, such as Alzheimer's disease, a notoriously difficult disease to diagnose based on clinical examination or radiology [23]. By using 2700 two-dimensional and 3795 three-dimensional images, an accuracy of over 85% could be achieved. Deep learning approaches for a variety of other CT/MRI applications have been described in recent publications showing good results for automated detection of pulmonary nodules/tuberculosis, postmortem hemorrhagic pericardium, and prostate cancer, amongst others [24–27].

Although encouraging, these results are likely only a modest estimate of the potential of CNNs in radiology in the future. These studies suggest that deep neural networks are very likely to consistently out-perform machine-learning models once large datasets (e.g. 100,000 cases) are available for training (Figure 4(B)). With international collaborations and consortiums, assembly of such large training image sets may become less of a bottleneck and it is expected that the generalizability and widespread use of CNNs in radiomics will grow. It will be important for large collaborations and consortia to achieve some degree of standardization and organization in compiled datasets to permit effective uptake for deep learning feature training. Supervised learning approaches, in particular, stand to benefit significantly from standardized labeling and data annotation systems. Both supervised and unsupervised approaches will benefit significantly from utilization of similar image generation and processing infrastructures, though some technical variation within data sets may be acceptable. An international initiative, comparable to that undertaken by the International Atomic Energy Agency (IAEA) for digital implementation of radiology, may assist in defining and implementing the standardization of these parameters for future deep learning image analysis approaches [28]. The development of truly effective image analysis tools in radiology stands to bring about transformational change in the way diagnostic radiology relates to patient care. Preliminary interpretation of scans could be performed almost immediately following image acquisition, permitting MRI and CT to move much closer to the point of care. In remote and underserved regions, which do not necessarily have access to expert radiologists, deep learning tools could provide invaluable diagnostic support in urgent clinical settings by
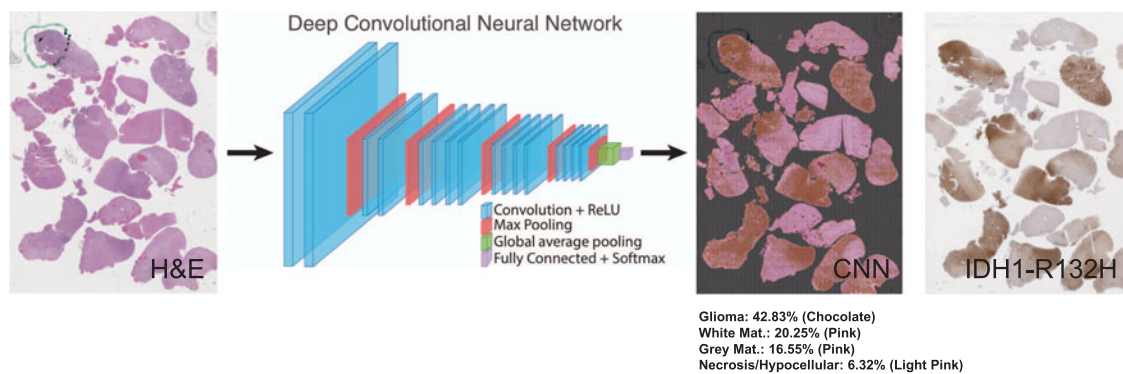
Glioma: 42.83% (Chocolate)
White Mat.: 20.25% (Pink)
Grey Mat.: 16.55% (Pink)
Necrosis/Hypocellular: 6.32% (Light Pink)

Figure 5. Visualization of histopathological image classification using CNNs. Class activation map (CAM) showing the localization of lesional tissue on a whole slide image (WSI) over 1 gigabyte in size, as detected by a trained CNN. This is a case of an oligodendroglioma, IDH-mutated, 1p19q co-deleted. For comparison, the immunostain for IDH1-R132H shows strong concordance with the "ground truth." Classes with the highest probability score can be used to provide annotations and a diagnosis for the entire WSI.

providing real-time image interpretation. In summary, while a diverse array of traditional machine learning approaches for computer vision already exist in the aforementioned cases of clinical practice, accumulating milestones in deep learning are showing that CNNs are emerging as the tool of choice to act as diagnostic aids for a diverse set of image-based pattern recognition tasks.

Deep learning for diagnostic surgical pathology

Traditional microscopic analysis of patient tissue by pathologists continues to serve as a useful tool to retrieve actionable diagnostic information in the acute setting. Intra-operative ("point of care") and emergent pathological analysis of patient tissue can be carried out using hematoxylin and eosin (H&E) stained slides [29] (Figure 1). Although effective, morphological analysis of tissue is a relatively qualitative art and a subspecialized skill. Moreover, some clinically relevant and immediately actionable questions are relatively challenging and laborious to resolve (e.g. margin assessment, identification of small metastatic foci, or precise tumor sub-classification) without immunohistochemical stains or molecular tests. Furthermore, other microscopic findings of prognostic significance (e.g. response to neoadjuvant treatment vs. radiation necrosis) may be independent of reproducible molecular changes [30]. Such information could find use in personalizing intraoperative surgical approaches or guide initiation of early precision therapies. With the increasing prevalence of digital scanners, whole slide imaging, and with the recent Food and Drug Administration (FDA) approval, pathologists are now also poised to benefit from automation of pathological analysis based on microscopy images (Figure 5). The most fundamental application of

CNNs in pathological analysis of histological images is the detection of small foci of residual cancer and metastasis [31]. Tumor cells exhibit notably different morphological features from normal cells on stained slides, especially in terms of cellularity, texture, color, and shape. A CNN method to detect presence of prostate and breast cancer metastasis in sentinel lymph nodes was recently validated, with AUC values of 0.99 and 0.88, respectively [32]. The model's high sensitivity allowed successful identification of all breast and prostate cancer metastases, and furthermore, excluded 30–40% of slides that contained benign and normal tissue. Workflow modification that help triage the most informative slides and highlight rare but actionable events (e.g. mitoses, positive tumor margins) aim to substantially improve the efficiency and accuracy of pathologists. CNNs can quickly provide direct binary classifications (presence or absence of tumor) on a gigapixel pathological slide, and localize the tumor for pathologist to review (AUC of 0.97) [33]. These experiences, however, also highlight some challenges. Although the classifier was quite sensitive in detecting slide images with metastases, it did not necessarily isolate all clusters of tumor cells on the slide and performed poorly, with false-positive results, when encountering artifacts. These limitations highlight the need for multiple visualization approaches and error reduction techniques that provide intuitive outputs for human pathologists to efficiently review computer generated whole slide interpretations. One popular approach is the use of "class activation maps (CAMs)" to superimpose and localize the coordination of detected abnormalities on a whole slide digital image for human review [34,35] (Figure 5). Other attractive applications of CNNs to pathology include objective differentiation of epithelial and stromal regions in colorectal cancer, cancer grading and accurate quantification of relatively

rare but important events indicating that the cancer is aggressive or benign, such as mitotic bodies and border of metastasis. This is another laborious process for pathologists that has now been conquered by deep learning [36–38]. Lastly, tumor type classification remains one of the key morphologic features that influences therapeutic options, including intra-operative personalized decision making (Figure 5). Classic cases include differentiating a glioma (treated with aggressive surgery) and other tumors (e.g. lymphoma) or inflammatory processes (e.g. multiple sclerosis and progressive multifocal leukoencephalopathy), from which patients benefit from medical, rather than surgical, management. Differentiating these entities in an intra-operative setting can be challenging, especially at hospitals where sub-specialized pathologists are not available. While most CNN models in histopathology have been applied to binary classification tasks, there is growing interest, including within our own group, in developing more complex multi-class classifiers [39–41]. It is likely that most autonomous diagnostic tasks will require a transition to CNNs that can carry out multiclass classifications of microscopic entities, a skill human pathologists currently excel at. These initiatives so far have shown to be much more challenging than binary classification tasks [40]. While a CNN was able to distinguish abnormal epithelial cells from normal cells with an accuracy of 98%, performance dropped significantly to 88% when the classifier was challenged to differentiate between five different subtypes of colon cancer (i.e. adenocarcinoma, mucinous carcinoma, serrated carcinoma, papillary carcinoma, and cribriform comedo-type adenocarcinoma). Intriguingly, the authors visualized feature maps of single neurons within the CNN and noted that the classifier learned more subtle features when differentiating colon cancer subtypes. Such feature map visualization techniques have remarkable potential for discovering biologically-relevant morphologic features exclusive to certain cancers.

Other visualization techniques aim at providing more transparent ways to visualize how CNNs store and organize learned information. For example, we recently showed the utility of applying the non-linear dimension-reduction technique, t-distributed Stochastic Neighbor Embedding (t-SNE) to pathological image analysis (Figure 6). This tool allows one to readily examine the inner mapping of histopathologic features by CNN and allows the classification process to be carried out in a more transparent way [41]. Importantly, by discretizing tissue types with non-neighboring decision boundaries, this visually interpretable classification approach helped reduce erroneous classifications when

unlearned subtypes and tumor classes were encountered (Figure 6). Such modifications that reduce errors are essential for translating image classifiers into the clinical and point of care setting.

Bridging computer-aided diagnostics (CAD) to clinical outcomes

Ultimately, objectifying the microscopic examination into a quantitative tool provides new opportunities to integrate multi-modal clinical information and define new and complex clinicopathologic patterns of clinical outcomes (e.g. survival rate and recurrent time). By adding a Cox proportional hazards layer to the end of a CNN trained on histologic images of gliomas, Mobadersany et al. show that they could define prognostic patterns that predicted progression time and survival [35]. Despite the complexity of time-to-event prediction, which requires clinical follow-up of a large cohort and differentiation of subtle features affecting risk, the model was able to provide a continuum of risk from histologic images. These models learned similar "high-risk" features routinely used by neuropathologists, including microvascular proliferation and necrosis and rivaled traditional prognostication criteria. These findings provide reassurance that CNNs and deep learning approaches can identify robust features with known biological significance, rather than only non-sensible abstract or artificial noise between images. Learning on these well-understood features provides optimism that these tools could generalize well to international centers when eventually implemented. Interestingly, they also discovered other subtler features of prognostic significance, including tumoral edema and invasiveness. Such features could provide more point of care feedback to surgeons and oncologists that could transform how gliomas and other tumors are managed in the acute setting. Importantly, when genomic information is available, these digital prognostic features could be integrated with molecular information and provide state-of-the-art predictions of survival. Similarly, predictions of clinical outcome for various other cancers have also been recently achieved by integrating CNN with other machine learning techniques, such as Recurrent Neural Network (RNN), Support Vector Machine, Logistic Regression, and Naïve Bayes [42]. The CNN (Long Short Term Memory (LSTM), a type of RNN) model can bypass trivial tasks such as tumor infiltrating immune cell quantification, mitotic figure counting, tissue entity identification, and tumor grading, and directly provide prognostic features that correlate with survival as strongly as Duke's stage and histologic grade
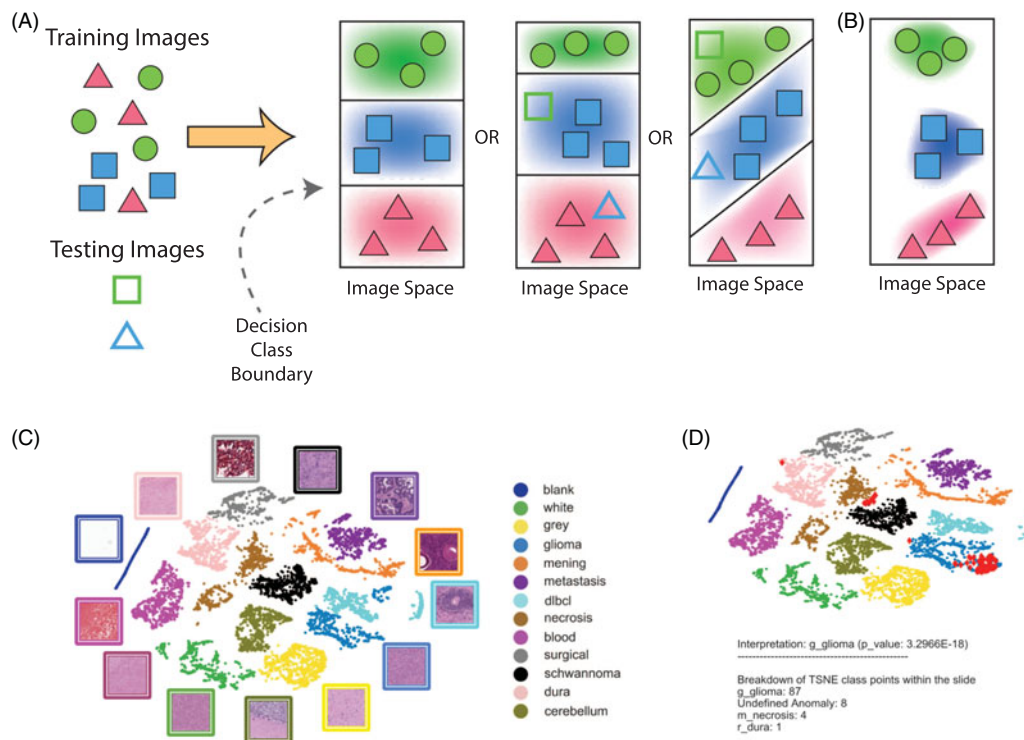
Figure 6. Visualizing CNN image classification. Neural networks are often described as "black boxes," as the features and weighting used for classification are not usually accessible to humans for review. (A) Hypothetical example of the image space stored within the CNN and the optimized classification boundaries for the solid (filled) objects. Since the image decision boundaries are unknown and often occupy larger areas than training examples, understanding classification errors can be challenging. (B) Visualization approaches (e.g. t-distributed stochastic neighbor embedding, t-SNE) that can help depict these image spaces could allow for more conservative decision boundaries to reduce errors and flagging challenging cases as undefined. (C,D) CNN learning (C) and classification of test slides by overlaying them on the t-SNE plot (red diamonds, Panel D) can be visualized using dimensionality reduction techniques. Collectively, these tools help provide transparency to CNN-based classifications and could be an important component to accelerate implementation of these tools in the clinic.

for colorectal cancer. Lastly, with the incorporation of genomics data (e.g. gene expression, methylation, and copy number alteration (CNA)) as well as proteomics data, some machine learning methods (e.g. Multiple kernel learning, Random Forest, and Boosting concordance index) can outperform Cox regression model in predicting survival times of breast cancer patients [43]. Overall, these studies highlight how synergism between deep learning, machine learning, and dimension reduction techniques on histologic images and molecular data could allow more readily available prognostication of cancer patients.

Regulatory considerations for clinical implementation

Recent advances in artificially intelligent approaches for real-world diagnostic settings have given rise to a plethora of ethical, regulatory, and legal considerations, many of which are incompletely resolved at this time. From an application development perspective, concerns revolve primarily around the legal and ethical implications of

using real patient data for training of deep learning systems. This issue was exemplified by a recent ruling in the United Kingdom (UK) by the Information Commissioner's Office (ICO) which noted that data from 1.6 million patients had been provided to Google's DeepMind without properly informing patients or obtaining their consent [44]. This highlights the importance of robust research and ethics board (REB) involvement in the development of AI tools in health care, to ensure that patients' rights and privacy are properly protected and respected. Even once platforms enter the implementation phase, data privacy concerns still remain, especially when data storage or analysis is performed by a third-party vendor outside of the patient's circle of care. Additional considerations for deep learning tools used in a clinical setting include validation of safety and efficacy targets, addressing medicolegal concerns surrounding medical errors which arise from use of these applications (i.e. misdiagnosis or incomplete diagnosis), and ensuring that tools are properly integrated into digital workflows in a manner compliant with regulatory standards.

Current regulatory frameworks for safe and ethical AI research and development in all industries, including healthcare, are almost non-existent. In October 2016, the National Science and Technology Council (NSTC) published a document entitled "The National Artificial Intelligence Research and Development Strategic Plan" in which the lack of regulatory frameworks for AI systems was acknowledged. While a need to "ensure the safety and security of AI systems" was described in this draft, little was offered in terms of formal regulatory standards or initiatives. In a bid to build some regulatory oversight, a bill entitled the "FUTURE of Artificial Intelligence Act of 2017" (Bill H.R(0).4625) has recently been introduced to U.S. Congress, and if passed could lead to the creation of a formal Federal Advisory Committee in the United States [45]. Regulatory frameworks in Canada and Europe remain similarly undefined, though member states of the European Union (EU) have recently signed to a Cooperation on Artificial Intelligence, which seeks to ensure implementation of legal and ethical frameworks, which protect privacy and personal data protection and maintain adequate transparency and accountability [46].

From a healthcare perspective, diagnostic platforms are not specifically regulated by the FDA, and are required only to adhere to regional data privacy rules and standards (such as those described in the EU Data Protection Directive and the U.S. Health Insurance Portability and Accountability Act (HIPAA)). The FDA is actively attempting to develop effective regulatory standards for diagnostic software platforms, including those which use deep learning architectures. A working model for development of regulatory oversight of these platforms, classified as "Software as a Medical Device" (SaMD) has been recently introduced by the FDA. This initiative, termed the Software Precertification Pilot Program, seeks to develop formalized assessment strategies to ensure that digital healthcare tools are delivered to patient care environments in a safe and effective manner [47]. Specifically, the program is attempting to pre-certify organizations involved in healthcare-related software creation based on the quality of their software design process and commitment to organizational excellence. While not yet a formal requirement for new AI tools, this effort may eventually lead to the implementation of a more robust and consistent regulatory standard for digital health platforms.

Discussion

The ability to generate and analyze large amounts of medical information has paved the way to a more precise and personalized approach to medical care. In many cases, we have rightly chosen to accept delays, in exchange for more accurate and refined molecular information. Technological improvements in image acquisition, image quality, and computational power, now allow for more objectivity and quantitative read-outs of visual medical data. The ability to practice with precision at the point of care could provide therapeutic, financial, and workforce improvements that could bring transformative changes in medicine. It is conceivable that primary care physicians and surgeons may be equipped with wearable cameras that can continually collect and immediately interpret the plethora of visual cues available in real-time at the point of care. Currently, most studies have focused on individual tasks. It is likely that developing integrated classification systems that incorporate multiple types of visual and non-visual information will help significantly accelerate the implementation and success of these technologies [35]. This has the potential to help triage costly molecular tests into their most appropriate and clinically actionable medical context.

Even with improvements in quality and size of available image databases, there are important challenges ahead. Like all other laboratory tests, standardization, quality assurance and benchmarks, as well as appropriate clinical context will need to be developed to reduce errors and improve patient safety. Automated image analysis is still in its infancy with many identified limitations [48]. Standardized validation sets that capture a comprehensive diversity of test images (including both routine and challenging cases) will need to be developed for each application. Similarly, rigorous criteria for approved equipment and output image quality and formats will also need to be explicitly defined to avoid extrapolation errors [48].

Other limitations of deep learning stem from its empirically derived nature that includes contributions from multiple programmers (e.g. different CNN architecture) and physicians (e.g. image annotators). This creates a significant challenge for liability in the event of computer-driven diagnostic errors. There is still debate as to who (e.g. image annotator, computer scientist, distributor, or the attending physician) will be held liable in the event of a diagnostic oversight. Situations where physicians have the authority to override computer-driven decisions will also need to be defined. This would be essential when new diseases arise that are not incorporated into the previous training sets. One solution is to build in mechanisms for computers to interpret and detect anomalies/novelty, and alert clinicians for careful review. This ability to detect and triage

challenging cases for human review has been extremely effective in allowing early implementation of machine learning tools in a variety of settings (e.g. using computers to automate bank check deposits). This is essential as most classifiers today are trained on relatively few diagnostic classes as compared to humans who can adapt and function well under diverse and evolving scenarios. Moreover, significant concerns arise from the relative obscurity ("black box") in how CNNs make classification decisions (Figure 6). Unlike human vision, which tends to identify objects based on global features (e.g. a human face), computers focus on varying combinations of relatively lower level features for classification (e.g. eye distance and skin color). The boundaries that determine the impact of these different features on the overall interpretation of the image, however, have been challenging to resolve. Although these differences lead to some impressive results in very controlled and closed environments, they make CNNs prone to errors that are difficult to predict *a priori*. Our group is addressing these limitations by developing visualization tools that help humans directly define more conservative decision boundaries and better understand why a computer reached a specific decision (Figure 6(B–D)). We believe that this and similar approaches to addressing anomalies may help provide much-needed transparency and better allow humans to more clearly understand computer errors, and promptly respond to them. Such information would allow physicians to remain as the primary decision makers and use AI as ancillary tools that provide additional information that needs to be integrated with clinical judgment.

It is important to also realize that for certain scenarios, precision from medical images may be impossible. Classifiers that can confidently communicate this information are still highly valuable to reduce over- and under-treatment, prior to more authoritative molecular or other ancillary testings. Similarly, cooperativity and integration of information across the available diagnostic modalities serve to synergize technological advancements.

The potential benefits and drawbacks of AI and neural networks in the acute care setting are still largely theoretical. Without systematic and comprehensive, longitudinal and practical experience, it is difficult to accurately predict the most significant concerns and areas of benefit. Practical experience and further technological improvements will likely need to arise and be addressed on a case by case basis. As with any new technology, especially those concerning vulnerable populations, it is most responsible to start with "low" and "slow" milestones. For the foreseeable future, it is likely that physicians will still be required to use their multi-faceted clinical knowledge and intuition as we move toward integrating pattern recognition and diagnostic decision-making support tools in the acute care setting.

Disclosure statement

The authors declare no conflicts of interests.

Funding

References

[1]   Abrahams E. Right drug-right patient-right time: personalized medicine coalition. Clin Transl Sci. 2008;1:11–12.

[2]   Abrahams E, Silver M. The case for personalized medicine. J Diabetes Sci Technol. 2009;3:680–684.

[3]   Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature. 2017;542:115–118.

[4]   Dolan NC. Skin cancer control in the primary care setting: are we making any progress? J Gen Intern Med. 2001;16:342–343.

[5]   Wilhelm M, Schlegl J, Hahne H, et al. Mass-spectrometry-based draft of the human proteome. Nature. 2014;509:582–587.

[6]   Kim MS, Pinto SM, Getnet D, et al. A draft map of the human proteome. Nature. 2014;509:575–581.

[7]   Djuric U, Rodrigues DC, Batruch I, et al. Spatiotemporal proteomic profiling of human cerebral development. Mol Cell Proteomics. 2017;16(9):1548–1562.

[8]   Bai H, Harmancı AS, Erson-Omay EZ, et al. Integrated genomic characterization of IDH1-mutant glioma malignant progression. Nat Genet. 2016;48:59–66.

[9]   Ceccarelli M, Barthel FP, Malta TM, et al. Molecular profiling reveals biologically discrete subsets and pathways of progression in diffuse glioma. Cell. 2016; 164:550–563.

[10]   Cancer Genome Atlas Research Network, et al. Comprehensive, integrative genomic analysis of diffuse lower-grade gliomas. N Engl J Med. 2015;372: 2481–2498.

[11]   LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521:436–444.

[12]   Deng J, Dong W, Socher R, et al. ImageNet: a large-scale hierarchical image database. Miami (FL): IEEE; 2009.

[13] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. New York (NY): Cornell University Library; 2014.

[14] Shen D, Wu G, Suk HI. Deep learning in medical image analysis. Annu Rev Biomed Eng. 2017;19: 221–248.

[15] Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA. 2016;316:2402–2410.

[16] Acharya UR, Oh SL, Hagiwara Y, et al. Deep convolutional neural network for the automated detection and diagnosis of seizure using EEG signals. Comput Biol Med. 2018;100:270–278.

[17] Rajpurkar P, Hannun AY, Haghpanahi M, et al. Cardiologist-level arrhythmia detection with convolutional neural networks. New York (NY): Cornell University Library; 2017.

[18] Vorravanpreecha N, Lertboonnum T, Rodjanadit R, et al. Studying down syndrome recognition probabilities in Thai children with de-identified computer-aided facial analysis. Am J Med Genet. 2018;176: 1935–1940.

[19] Knaus A, Pantel JT, Pendziwiat M, et al. Characterization of glycosylphosphatidylinositol biosynthesis defects by clinical features, flow cytometry, and automated image analysis. Genome Med. 2018; 10:3.

[20] Komeda Y, Handa H, Watanabe T, et al. Computer-aided diagnosis based on convolutional neural network system for colorectal polyp classification: preliminary experience. Oncology. 2017;93:30–34.

[21] Rajpurkar P, Irvin J, Zhu K, et al. CheXNet: radiologist-level pneumonia detection on chest x-rays with deep learning. New York (NY): Cornell University Library; 2017.

[22] Arevalo J, González FA, Ramos-Pollán R, et al. Representation learning for mammography mass lesion classification with convolutional neural networks. Comput Methods Programs Biomed. 2016;127: 248–257.

[23] Gao XW, Hui R, Tian Z. Classification of CT brain images based on deep learning networks. Comput Methods Programs Biomed. 2017;138:49–56.

[24] Lakhani P, Sundaram B. Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks. Radiology. 2017;284:574–582.

[25] Ebert LC, Heimer J, Schweitzer W, et al. Automatic detection of hemorrhagic pericardial effusion on PMCT using deep learning – a feasibility study. Forensic Sci Med Pathol. 2017;13:426–431.

[26] Hua KL, Hsu CH, Hidayati SC, et al. Computer-aided classification of lung nodules on computed tomography images via deep learning technique. Onco Targets Ther. 2015;8:2015–2022.

[27] Liao S, Gao Y, Oto A, et al. Representation learning: a unified deep learning framework for automatic prostate MR segmentation. Med Image Comput Comput Assist Interv. 2013;16:254–261.

[28] IAEA. Worldwide implementation of digital imaging in radiology. Vienna, Austria: IAEA; 2015.

[29] Djuric U, Zadeh G, Aldape K, et al. Precision histology: how deep learning is poised to revitalize histomorphology for personalized cancer care. NPJ Precis Oncol. 2017;1:22.

[30] Chui MH, Kandel RA, Wong M, et al. Histopathologic features of prognostic significance in high-grade osteosarcoma. Arch Pathol Lab Med. 2016;140: 1231–1242.

[31] Wang D, Khosla A, Gargeya R, et al. Deep learning for identifying metastatic breast cancer. New York (NY): Cornell University Library; 2016.

[32] Litjens G, Sánchez CI, Timofeeva N, et al. Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. Sci Rep. 2016;6:26286.

[33] Liu Y, Gadepalli K, Norouzi M, et al. Detecting cancer metastases on gigapixel pathology images. New York (NY): Cornell University Library; 2017.

[34] Faust K, Xie Q, Han D, et al. Visualizing histopathologic deep learning classification and anomaly detection using nonlinear feature space dimensionality reduction. BMC Bioinformatics. 2018;19:173.

[35] Mobadersany P, Yousefi S, Amgad M, et al. Predicting cancer outcomes from histology and genomics using convolutional networks. Proc Natl Acad Sci USA. 2018; 115:E2970–E2979.

[36] Barker J, Hoogi A, Depeursinge A, et al. Automated classification of brain tumor type in whole-slide digital pathology images using local representative tiles. Med Image Anal. 2016;30:60–71.

[37] Cireşan DC, Giusti A, Gambardella LM, et al. Mitosis detection in breast cancer histology images with deep neural networks. Med Image Comput Comput Assist Interv. 2013;16:411–418.

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

[39] Haj-Hassan H, Chaddad A, Harkouss Y, et al. Classifications of multispectral colorectal cancer tissues using convolution neural network. J Pathol Inform. 2017;8:1.

[40] Xu Y, Jia Z, Wang LB, et al. Large scale tissue histopathology image classification, segmentation, and visualization via deep convolutional activation features. BMC Bioinformatics. 2017;18:281.

[41] Faust K, Xie Q, Han D, et al. Visualizing histopathologic deep learning classification and anomaly detection using nonlinear feature space dimensionality reduction. BMC Bioinformatics. 2018;19:173.

[42] Bychkov D, Linder N, Turkki R, et al. Deep learning based tissue analysis predicts outcome in colorectal cancer. Sci Rep. 2018;8:3395.

[43] Sun D, Li A, Tang B, et al. Integrating genomic data and pathological images to effectively predict breast cancer clinical outcome. Comput Methods Programs Biomed. 2018;161:45–53.

[44] Powles J, Hodson H. Google deepmind and healthcare in an age of algorithms. Health Technol (Berl). 2017;7: 351–367.

[45] Bagloee SA, Tavana M, Asadi M, et al. Autonomous vehicles: challenges, opportunities, and future

implications for transportation policies. J Mod Tansp. 2016;24:284–303.

[46] EU Declaration on Cooperation on Artificial Intelligence – European Commission. [cited 2018 Sep 24]. Available from: https://ec.europa.eu/jrc/communities/community/digitranscope-digital-transformation-and-governance-human-society/document/eu-declaration

[47] Patel Bakul. Developing a software precertification program: a working model. Silver Spring (MD): Food and Drug Administration; 2018.

[48] Lambin P, Leijenaar RTH, Deist TM, et al. Radiomics: the bridge between medical imaging and personalized medicine. Nat Rev Clin Oncol. 2017;14: 749–762.