nature medicine

REVIEW ARTICLE

https://doi.org/10.1038/s41591-021-01343-4

Deep learning in histopathology: the path to the clinic

Jeroen van der Laak[1,2]✉, Geert Litjens[1] and Francesco Ciompi[1]

Machine learning techniques have great potential to improve medical diagnostics, offering ways to improve accuracy, reproducibility and speed, and to ease workloads for clinicians. In the field of histopathology, deep learning algorithms have been developed that perform similarly to trained pathologists for tasks such as tumor detection and grading. However, despite these promising results, very few algorithms have reached clinical implementation, challenging the balance between hope and hype for these new techniques. This Review provides an overview of the current state of the field, as well as describing the challenges that still need to be addressed before artificial intelligence in histopathology can achieve clinical value.

Breakthroughs in the field of artificial intelligence (AI) have had a major impact on society worldwide in the past 5 years. In the field of medicine, and more specifically in diagnostic disciplines (for example, radiology and pathology), initial results from the application of AI to patient data are very promising[1–3]. Diagnostic disciplines often rely heavily on the recognition of patterns in data, such as images, by physicians and the interpretation of such patterns in the wider context of the patient. However, it was shown that for many diagnostic tasks, reproducibility among physicians is less than optimal[4,5]. Also, as a result of the increase in treatment options, more accurate diagnostics are needed to meet the requirements of precision medicine, which may exceed the capabilities of human visual inspection[6]. Detecting and accurately quantifying patterns in medical data using AI could therefore aid diagnostic processes, making them more efficient and reproducible, and increasing accuracy and precision.

An area in which the use of AI is particularly appealing is in the analysis of histopathological tissue sections, which currently requires specialized doctors, pathologists, to carefully assess (sometimes large numbers of) gigapixel-sized images. Pathologists diagnose and grade diseases such as cancer and inflammatory diseases, based on a variety of tissue features (for example, disturbed tissue architecture, the presence or absence of specific cell characteristics or the presence of, for example, an abundance of inflammatory cells). While there is a worsening shortage of pathologists, their workload is increasing as a result of larger numbers of cases, and the requirement for more extensive diagnoses to identify the most optimal treatment for patients.

Using AI to analyze tissue sections is often referred to as computational pathology (CPATH)[7,8] (see Box 1), the current applications of which rely heavily on the use of deep neural networks (so-called deep learning). Research in this area began as early as the 1960s, with the initial application of image analysis algorithms to images of cells. Individual cells in blood smears could be classified into subtypes, on the basis of quantitative cell characteristics such as size, shape and chromatin distribution, to analyze the blood composition and help diagnose a range of diseases[9]. Early CPATH applications attempted to implement computational features that were painstakingly matched to a biological process or shape, and were later replaced by radiomics or pathomics approaches using generic feature banks of texture descriptors (that is, a quantitative description of the characteristics of image textures, such as orientation, contrast and so on)[10], operating under the assumption that complex classifiers could eventually find the intricate relationships among these features for specific classification tasks (e.g., ref. [11]). For example, Kather et al. showed that combining five different types of texture descriptors resulted in a classifier that could recognize tumor and stroma in colorectal tissue sections with 98.6% accuracy[12].

The almost complete transition from feature engineering to deep learning occurred for several reasons. For medical imaging, and thus also for CPATH, perhaps the most important reason is the fact that the construction of algorithms (almost) entirely by training, rather than by explicit programming or by using predefined filters, yields powerful, hierarchical feature representations that, in most cases, outperform more traditional image analysis methods[3]. As a consequence, the need to have domain knowledge to achieve good results is reduced because feature engineering requires the definition of problem-specific features, whereas in deep learning the networks learn meaningful features autonomously from the data. Automatically learning features from the data also leads to reduced implementation time. In feature engineering, crafting meaningful characteristics for the data at hand generally requires several iterations per feature and sometimes lengthy and repeated discussions with pathologists to understand what cues are used during their diagnostic process. In the era of deep learning, such trajectories can be reduced to months, sometimes even weeks, while breaking boundaries in terms of diagnostic performance. Lastly, the importance of freely available source code for the most successful neural network architectures cannot be understated.

The history of CPATH before deep learning has been reviewed elsewhere[13], as has the application of deep learning for histopathology from a technical perspective[14]. Therefore, in this Review we provide an application-based approach to the use of deep learning in histopathology with an emphasis on clinical value, future developments and challenges still to overcome before true patient value is achieved, as well as briefly retracing the main milestones in CPATH from the past decade. We have limited the discussion to studies in which computational analysis is applied to digitized bright-field microscopic tissue sections in combination with selected metadata, as these form the vast majority of current research.

[1]Department of Pathology, Radboud University Medical Center, Nijmegen, the Netherlands. [2]Center for Medical Image Science and Visualization, Linköping University, Linköping, Sweden. ✉e-mail: jeroen.vanderlaak@radboudumc.nl

Box 1 | Definitions

Deep learning
A machine learning approach in which algorithms are trained for a specific task (or set of tasks) by exposing a multilayered artificial neural network to (typically a large amount of) training data, without the need for handcrafted engineering of features to be extracted from the data. The resulting algorithm has learned a hierarchical representation of the data that is subsequently used for tasks such as classification, detection or segmentation. The term deep refers to artificial neural networks built using many layers, in other words a deep neural network.

Digital pathology
The digitization of the traditional diagnostic process of analyzing cells and tissue with a microscope via whole-slide scanners and computer screens.

Computational pathology
The computational analysis of digital images obtained through scanning slides of cells and tissues.

Radiomics/pathomics
Techniques to extract a (usually very large) set of features from radiological or histopathological digital images, respectively, using computational algorithms of data analysis. These features are successively used to feed (usually supervised) prediction models targeting clinically relevant end points, such as prognosis.

End-to-end training
In the context of machine learning models, possibly consisting of a pipeline with multiple steps, end-to-end training refers to the procedure of learning the optimal value of all parameters of a model simultaneously rather than sequentially (that is, one step at a time).

Whole-slide images
Digital images obtained by digitizing complete histopathological glass slides using a high-resolution scanner.

Convolutional neural networks
Deep learning approach consisting of a series of convolutional layers to process data (usually bi-dimensional) from input to output. Each layer implements the convolution operation between the input data and a set of filters (that is, small matrices), whose numerical values are automatically learned in an end-to-end training fashion.

Graphics processing units
Microprocessor specifically designed to process many data samples simultaneously, such as parts of digital images or features extracted from images.

Image segmentation
The operation of decomposing the semantic content of an image into multiple segments, where each segment contains pixels belonging to the same semantic category (for example, the tumor region).

U-Net models
Deep learning models based on two convolutional neural networks, one that encodes the input image into a set of features, and one that decodes those features to produce a segmentation output. The name, introduced in 2015 by Ronneberger et al.[145], indicates the U shape that the two convolutional neural networks form, where the encoder and decoder are connected via skip connections.

Data augmentation
The operation of artificially modifying some properties of input data (for example, image contrast, orientation, color and so on) with the aim of feeding a computational model with multiple variations of the same piece of data.

Model regularization
In machine learning, indicates the process of constraining a model's parameters to small values, discouraging complex models, therefore reducing the risk of overfitting the training data.

Trends in CPATH

CPATH has moved forward substantially in the past 10 years as a result of strong improvements in microscopic scanning devices, which enable the acquisition of whole-slide images (WSIs), progress in the development and decreases in the cost of computing hardware, and advances in AI. The field has followed a trend that was previously experienced by the computer vision community, which focuses its research on computational analysis of natural images (that is, real-world photographs and videos). Initial reports in 2011 of efficiently training convolutional neural networks (CNNs), a particular type of deep learning algorithm, using graphics processing units (GPUs)[15], led to the design of much deeper CNNs that outperformed the state of the art (mostly using machine learning based on handcrafted features) in the classification of natural images. Specifically, in the ImageNet challenge, which required categorization of one million photographs in a thousand different classes ranging from specific breeds of dogs to airplanes and cars, these new deep neural networks reduced the error rate from 25% to 4% in only 3 years. CPATH researchers took note of the successes of applying CNNs in computer vision, initially presenting methods that solely focused on the analysis of small cropped areas from WSIs, such as mitosis counting[16]. Methods that used entire WSIs followed for applications such as breast cancer segmentation[17], glioma classification[18], non-alcoholic fatty liver disease[19], assessment of renal transplant biopsies[20] and prostate cancer detection[21].

In parallel with the progression to more advanced AI models in histopathology, the complexity of the tasks to be solved and the size of publicly available datasets began to grow. In 2016, the CAMELYON challenge was proposed with the aim of developing CPATH solutions for the detection of breast cancer metastases in sentinel lymph nodes[22]. The introduction of the CAMELYON dataset was a game changer in the field of CPATH, as it made available for the first time the largest collection ($n = 1{,}399$) of fully manually annotated WSIs of sentinel lymph nodes of patients with breast cancer. Participants in the challenge had to solve two tasks designed to mimic routine tasks in pathology diagnostics: finding tumor regions in each lymph node and consequently predicting the presence of tumors at a WSI level. The impact of CAMELYON was similar in magnitude to that of ImageNet on the computer vision community. The large set of data and clinical focus of CAMELYON stimulated the creativity of both researchers and industry, who pushed forward the development of AI for metastasis detection, thereby enabling CPATH methods to make a leap from both academic and commercial technological perspectives. Furthermore, CAMELYON also attracted machine learning powerhouses such as Google to the field of CPATH[23],

contributed to establishment of several CPATH start-up companies and influenced government policy in the USA[24]. Today, many research papers are published on CPATH developments, focusing on a whole array of clinical applications. These applications are often found to approach or even surpass the performance of pathologists for specific tasks. To facilitate the development of such applications, ever larger datasets with associated annotations are required, posing challenges in terms of data collection and annotation production.

CPATH for clinical practice. The CAMELYON challenge provided a stimulus for researchers and industry to focus on the actual impact of CPATH applications in pathology clinical practice. Current applications include tumor detection and classification (often by subtype[23,25–39]), image segmentation[40–50], cell detection and counting[51–55], mitosis detection[56–60], analysis of kidney transplant biopsies[20] and tumor grading[61–63] among others. An example of a CPATH application for automatic tissue segmentation using a combination of U-Net models[20], as well as the corresponding ground truth, is shown in Fig. 1. Figure 1a shows a zoomed-in region of a periodic acid–Schiff-stained kidney biopsy, in which glomeruli, tubuli, capillaries and so on can be recognized. In Fig. 1b, the expert annotation is shown, which is used to validate the output of the CPATH solution (purple, glomeruli; blue, proximal tubuli; orange, distal tubuli; green, atrophic tubuli; and so on). Figure 1c shows the output of the CPATH models, which clearly corresponds very well with the human annotation.

In the context of clinical practice, automating repetitive and time-consuming tasks such as the analysis of tissue samples obtained by biopsy and excised lymph nodes can have a tremendous impact on the optimization of the clinical workload of pathologists. Tissue samples from the breast, colon and cervix are taken in large numbers as a consequence of population screening programs, and large numbers of lymph nodes per patient are resected during surgery, resulting in large numbers of (mostly negative) slides to be checked by pathologists. In these situations, AI algorithms could flag suspicious regions or slides for inspection or, in the future, assess cases autonomously.

In addition to automating current diagnostic tasks, CPATH methods can also be used to support pathologists with additional information; for example, by showing the 2-mm^2 hotspots of mitotic cells in breast cancer WSIs that are required for tumor grading as advised by guidelines for treatment of patients with breast cancer (for example, as published by the American Society of Clinical Oncology)[64]. This approach performs similarly to pathologists and can reduce inter-observer variability[65]. Highlighting regions of prostate cancer using different colors to represent different Gleason grades[4,66] and highlighting lung cancer growth patterns by adenocarcinoma subtype[38,67] using CPATH methods have produced similar results. Furthermore, a combination of segmentation, detection and classification methods can enable the objective quantification of established biomarkers that are used in clinical practice. One example is the assessment of tumor-infiltrating lymphocytes[68], which can be achieved by segmenting stromal regions of a slide and detecting intrastromal lymphocytes by hematoxylin and eosin (H&E) staining[53,69] or by immunohistochemistry (IHC)[52]. Using this method, the presence of tumor-infiltrating lymphocytes was shown to correlate with recurrence and genetic mutations in lung adenocarcinoma[70]. Other examples of biomarkers include those related to the amount of intratumoral stroma[71], such as the tumor–stroma ratio[72], which can be assessed by computing the ratio between the tumor and tumor-associated stroma obtained via image segmentation, and the quantification of programmed death-ligand 1 (PD-L1)-positive cells, which is used to stratify patients for immunotherapy and can be achieved via the detection of positive (and possibly negative) cells, by segmentation of PD-L1-positive and PD-L1-negative regions[73] or even predicted from H&E slides[74].

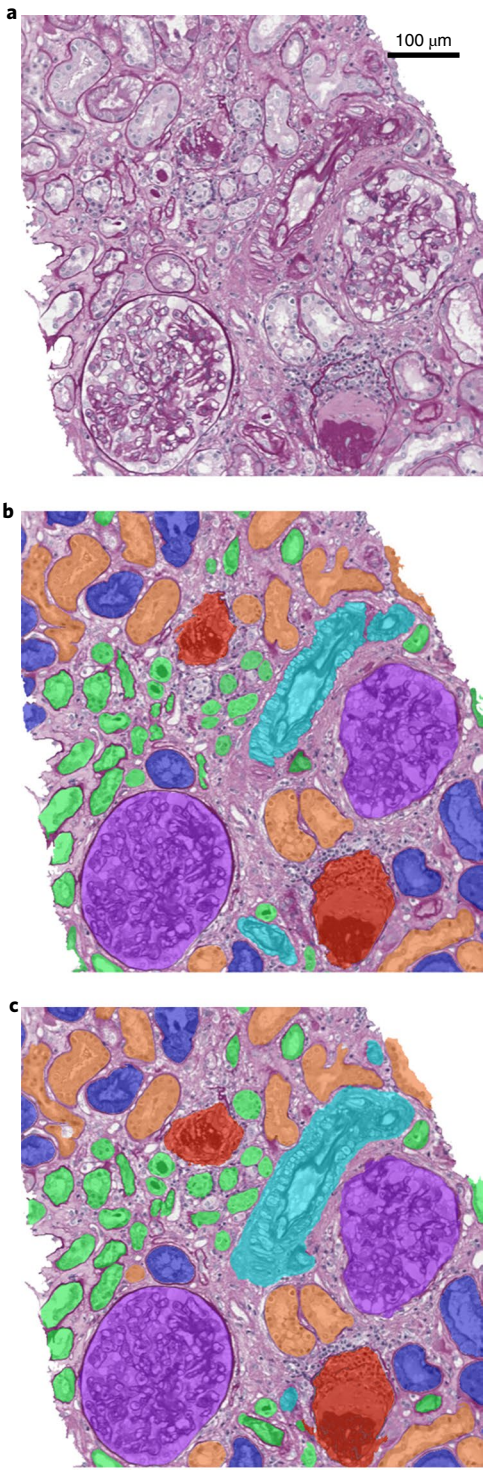

Fig. 1 | CPATH for tissue segmentation. A CPATH algorithm for kidney tissue segmentation based on CNNs has been applied to a periodic acid–Schiff-stained section of kidney tissue. **a**, Original image. **b**, Original image that has been manually annotated by an expert. **c**, The result of processing the original image using a deep learning algorithm[20,144]. No added color indicates interstitium; purple indicates glomerulus; red indicates sclerotic glomerulus; dark blue indicates proximal tubule; orange indicates distal tubule; green indicates atrophic tubule; and turquoise indicates artery or arteriole.

Large-scale datasets. Following the promising results of the early CPATH applications, the size of datasets has increased, leading to an increasing number of multicentric efforts to cope with the large

variability in staining, image quality, scanning characteristics and tissue preparation across different laboratories. One example of how datasets have grown is in the use of AI for prostate cancer detection, one method of which was developed using a dataset of 254 prostate tissue samples in 2016[21], whereas a method proposed in 2019 used a dataset with >24,000 prostate tissue samples[75]. As the scale of datasets grew, CPATH methods began to approach and even surpass the performance of pathologists[4,5,76].

However, although collecting a large number of WSIs is a manageable task for pathology laboratories and medical centers, collecting annotations remains an obstruction to the scaling of CPATH algorithms. Annotations can mean both the manual annotation of image regions (such as the identification of regions of tissue or the location of specific cells types) and clinical annotations (such as assessment of molecular subtypes, treatment response and survival). Acquiring manual annotations of images is a tedious task that requires domain expertise and is typically performed by (resident) pathologists. By contrast, clinical annotations require access to pathology reports and electronic patient records, either from a hospital (to retrieve information about grades, molecular subtypes or treatment responses) or from a regional or national registry (to retrieve information about survival), and can be provided only by authorized clinical researchers or data managers. Clinical annotation of WSIs tends to be easier to achieve than manual annotation, and has resulted in large datasets in several studies (for example, for prostate cancer[75], lung cancer[26] and colorectal cancer[77]). Still, building CPATH models using only clinical annotations will not be possible or efficient for every application in histopathology. For instance, if the feature that is critical for arriving at a certain diagnosis is present only in very small regions of the WSI, it may require a very large number of cases before the CPATH model has learned to perform the task. Therefore, manual annotations will still be needed, necessitating development of techniques to facilitate efficient production of these annotations.

Several approaches have been proposed to address the need for manual annotations in large-scale datasets. One straightforward approach is to simply scale the number of annotators with the data by involving a large number of experts. This approach has the advantage of guaranteeing high-quality annotations, but it is very expensive as it involves a fairly large number of experienced physicians. This solution was adopted by a study in 2020[78], in which 12 senior pathologists were involved in exhaustively manually annotating >2,000 WSIs of gastric cancer; the agreement among all experts involved was used as a reference standard. A similar approach was used in the TUPAC challenge in 2016 to define a reference standard for mitosis detection by combining the opinions of a panel of pathologists[56]. An alternative approach to using pathologists is to assign manual annotations to a set of people with different amounts of expertise, ranging from medical students to junior and senior pathologists[53,79,80]. In previous studies that used this approach, manual annotations were crowdsourced using web-based platforms such as Mechanical Turk[79,80]. However, in all cases, manual annotations were finally reviewed and approved by (resident) pathologists.

Staining techniques such as IHC, in which antibodies can be used to target specific types of tissue or cells, may also provide valuable support to manual annotations. This strategy was used to make manual annotations of breast cancer metastases to lymph nodes in the CAMELYON challenge, in which two serial sections were stained with cytokeratin (CK) and H&E, and CK was used to guide the manual annotation procedure[22]. This approach has the advantage of providing strong supervision to the annotator and avoiding false negatives and false positives in the annotated reference standard. Another useful technique is restaining, which provides an alternative to serial sections and enables the same slide to be subsequently stained with, for example, H&E and IHC, and the two digitized slides to be aligned via registration algorithms. This technique guarantees that exactly the same cells and tissue compartments are present in both slides, and that the positive marker in IHC can be transferred to H&E, de facto producing a strong basis for making accurate annotations automatically. This approach has been adopted for the detection of mitotic figures using phosphohistone H3 as a reference standard[57], for the segmentation of prostate epithelium using CK as a reference standard[81] and for the detection of epithelial cells in breast cancer using CK and Ki67 (ref. [82]). Restaining techniques enable the number of cases to be scaled at relatively low cost and with only a minimal interaction from human experts, thus reducing variability due to inter-observer disagreement, which is a well-known limitation in applications such as the detection of mitotic figures[57].

Weakly supervised learning. Another approach to reduce the burden of manual annotations is to consider CPATH algorithms that are trained in a weakly supervised fashion. In the context of image segmentation, weak supervision can come in the form of sparse manual annotations (for example, annotation of only small regions using dots or scribbles, as opposed to full supervision via dense annotations, in which all pixels of the image are manually labeled)[83,84]. Several groups have shown that weak supervision combined with advanced learning strategies in model development can approach the performance of fully supervised systems, particularly when sparse and dense annotations are combined. On the basis of this idea, weak supervision has been used to address several segmentation and detection problems in CPATH methods[43,50,60,85–87].

In weakly supervised WSI classification (for example, making a single prediction for the entire WSI), only a single label per image is available for model development, and methods based on manual annotations are no longer applicable. This setting is appealing in terms of scalability because information contained in clinical annotations is often sufficient to define the image-level target (such as the presence of cancer in WSIs) without the need to make manual annotations of cancer regions. Furthermore, clinical annotations can often be extracted from pathology reports and health records[88,89], opening a new avenue for automated analysis of those reports and for extraction of labels, with the potential to scale up to several thousand cases, which would be impossible to manually annotate. As an example, this type of challenge was proposed as one of the tasks in the TUPAC competition in 2016, in which participants were asked to predict a proliferation score derived from clinical annotations, such as molecular tests, for WSIs of breast cancer, which is impossible to manually delineate with annotations in the WSIs[56].

Technically speaking, WSI classification would not be different from the image classification performed in computer vision, in which CNNs are trained end-to-end to predict the presence of categories in natural images using image-level labels. However, end-to-end approaches cannot straightforwardly be applied to WSI classification, mainly because gigapixel WSIs are too large in size and do not fit into the memory of modern GPUs. Even switching to central processing unit computation would not resolve this problem, as a single WSI can easily require tens of gigabytes of memory at full resolution. Researchers have tried to overcome this limitation through different methodological innovations. A simple approach to tackle this problem is to assume that all patches in the WSI contain morphological information that correlates with the WSI-level label; for example, all patches extracted from a WSI that contains tumor also contain tumor. Despite the simplicity of this assumption, it can be effective for some applications[26,90], although it will not work when rare or small objects have to be found[91], such as small metastases in lymph nodes. The previous assumption can be refined by adopting a multiple instance learning approach[92], in which at least one small region in the image is considered to contain morphological information that is needed to classify the image; for example, the presence of a single small region containing cancer is sufficient to label the entire WSI as containing cancer[75,93,94].

Another approach to make end-to-end training possible using WSIs is to directly address the large WSI size as the main limitation, with the aim being to make the input size smaller so that the WSI can be processed by modern hardware. Recent approaches based on this idea rely on WSI compression using neural networks[95] under the assumption that semantic information can be kept in the compressed version of the entire WSI, which can then be used for downstream classification tasks while reducing the data size. Other approaches decompose the end-to-end training procedure using WSIs into parts and use advanced engineering techniques known as gradient checkpointing to temporarily store intermediate results[96]. These approaches make use of modern GPUs to train with very large input sizes, with the aim of scaling up to the use of entire WSIs as input.

In recent years, a number of CPATH methods have been presented (some of which use end-to-end learning) to further enhance the performance of the pathologist by providing information currently impossible to capture by sole visual inspection of histopathology slides, such as prediction of response to chemotherapy or immunotherapy, or even future events such as recurrence or survival[97–104] as well as the presence of genetic mutations[95,105–110] or molecular subtypes[30,111,112]. These CPATH techniques could have a role in the discovery of predictive and prognostic biomarkers, as well as potentially being used to understand tumor growth mechanisms.

Current challenges

Although considerable progress has been made in CPATH in the past 5 years, both in terms of algorithm performance and the development of novel methodologies, many challenges still exist. Some of these challenges, such as the lack of public datasets that are truly representative of clinical practice, stand in the way of true clinical adoption of CPATH algorithms. Other challenges such as difficulties in explaining how CPATH algorithms work are, in our view, less of a barrier than often thought[113]. In this section, we highlight some important challenges and the work that has already been done to tackle these issues.

Generalizability of CPATH algorithms to clinical practice. Although dataset sizes for developing CPATH algorithms have grown substantially over the past few years, many still lack an important characteristic in that they are not representative of the type of data that is encountered in clinical practice[114]. These data have many more sources of variation than the datasets used in research papers. Although most work now tries to account for variations caused by different scanners or staining techniques by including data from different laboratories, the number of laboratories included is typically too small for a true assessment of generalizability[115]. The number of laboratories that would need to be included to be representative would depend on the diagnostic question and, up until now, this aspect of CPATH has been poorly investigated. Other sources of variation have not yet been taken into account in CPATH, such as differing patient populations between centers or countries, although they are starting to be considered in other fields such as radiology[116]. Such variation can cause subtle sources of bias in CPATH algorithms, as seen in other settings[117].

These generalization issues are highlighted by the well-known phenomenon of CPATH algorithms performing optimally on data from the source(s) they were trained on, but performing (sometimes much) less well on data from other sources. For example, the application of a trained model for detection of prostate cancer to WSIs taken from the same dataset that was used for algorithm construction, but that had been rescanned on a different WSI scanner, gave an area under the curve reduction of 2.65%, whereas model performance dropped by 5.84% when applied to WSIs from an external dataset[75]. Examples of performance drops in the presence of external test data can also be also found in several other studies[61,118,119]. Limited generalizability of algorithms is probably the single most important obstacle for wide-scale implementation of CPATH techniques in the clinic.

To make CPATH algorithms as robust as possible in response to variations that are likely to be encountered in real-world practice, it is pivotal to establish a training set that contains as much variation as possible, including data from different staining batches, scanners and medical centers. Additional (artificial) variability may be introduced by data augmentation techniques, with a particular focus on color augmentation, to mimic differences in staining from different pathology laboratories[120]: image patches can be transformed before being used in the training process by applying random rotation, flipping, addition of noise, blurring and color shifting.

An alternative (or possibly complementary) approach to deal with variations between data sources is the normalization of images to a common standard[118,121–124]. The hypothesis is that if variability can be removed and all (future) target images can be translated to a well-defined standard (mostly in terms of color specifications), then even a CPATH algorithm built on a narrow set of training images would perform consistently well. The price that is paid for this approach is the need to transform every target image before applying the CPATH model, which may be computationally costly. Both data augmentation and image normalization are necessary requisites to increase generalizability of deep learning models[120], and should therefore be considered in the development of any CPATH approach.

Another important issue to contend with is the fact that a CPATH algorithm will recognize only the patterns it was trained to recognize. For example, if an algorithm that was trained to detect breast cancer metastases in lymph nodes was confronted with lymphoma, the outcome would be uncertain. If such an algorithm were used to filter out obviously negative WSIs (without lymph node metastases), which do not need to be inspected by a pathologist, serious diseases could remain undetected. One possible solution is to train algorithms for all possible pathologies; however, such an approach may be impractical or even impossible in most cases. An alternative approach is the development of CNN techniques that yield, in addition to the network output, a score that expresses the certainty of the CNN for that specific output[125,126], which would offer CNNs the ability to essentially state 'I do not know'.

Validation of CPATH algorithms. Algorithm validation is crucial to understand the usefulness of CPATH algorithms for broad applications and to collect evidence on the safety and accuracy of algorithms for regulatory approval. Different levels of validation can be used during algorithm development (Fig. 2). Typically, CPATH algorithms are validated in multiple ways during development. As part of the actual algorithm construction, the training process is monitored using a set of cases that are held apart from the rest of the dataset and are therefore not used for model training (often referred to as the validation set, which is usually relatively small). Deviations between the results obtained with the training data and the validation set may indicate overtraining and suggest that further action is required (for example, use additional techniques such as data augmentation or model regularization, or reduce the complexity of the deep learning architecture).

Many CPATH studies use a fully independent set of cases (a test set) to subsequently assess performance of the final model. In most studies, these are from the same data source (so-called internal validation) and as such have characteristics that are very similar to the cases used for training. If the training dataset is of limited size, sometimes cross-validation is used rather than applying fully independent hold-out sets. In cross-validation, multiple models are trained with different non-overlapping subsets of cases for testing and training, and an average performance score is given. Using cases that were not used for model training but were held separately from the rest of the dataset for performance assessment is good practice to arrive at a first indication of how well the algorithm works, but

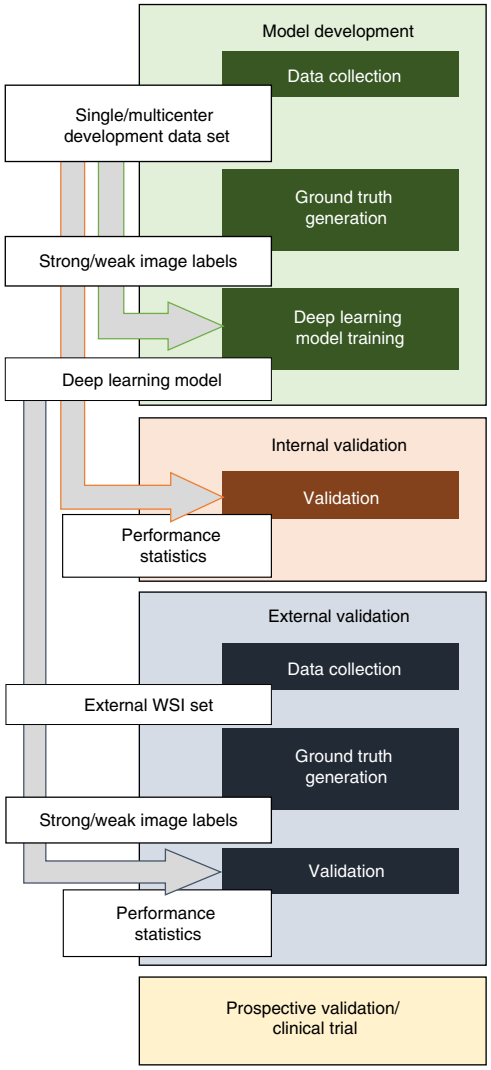

Fig. 2 | Validation of CPATH algorithms. A schematic representation of the main steps involved in the development and validation of CPATH algorithms using deep learning. Different levels of validation are shown in colored boxes vertically (validation during training, internal and external validation, and prospective validation/clinical trial). The consecutive activities in a typical deep learning workflow are shown in darker colored boxes, with the corresponding products of these actions shown in white boxes on the left. Arrows indicate which products are inputs for the activities.

should be regarded as only a first step towards a realistic assessment of the usefulness in clinical practice[127].

A next step, which has been used in several studies[4,71,75,103], is to validate the CPATH algorithm using an entirely separate set of cases from a source that was not included in the training data (known as external validation; Fig. 2). Such validation gives an indication of how well the algorithm performs in a new diagnostic situation, and can uncover problems with generalizability[128,129]. The availability of publicly accessible benchmark datasets[5,56] may be very helpful for this purpose, as it allows fair comparison between different CPATH algorithms[115]. Such datasets may also support regulatory approval[130]. However, even good performance on an external dataset is not proof of the clinical usefulness of algorithms, and should not be regarded or reported as such[131]. Some of the hype around the promises of AI in the medical domain may in fact result from the overly optimistic interpretation of results of external validation studies. As with any innovation in health care, well-conducted prospective studies are required to provide the evidence necessary to truly understand the added value of CPATH deep learning algorithms and pave the way for clinical implementation[115,131].

With the increasing autonomy of CPATH solutions, more rigorous clinical validation and regulatory approval[132] will also be required. Techniques that potentially influence diagnostic decision-making (rather than aiming to only increase efficiency) may need to be investigated in randomized clinical trials, which are currently still very rare for AI applications[133]. Ideally, such trials would use clinical outcomes as end points to demonstrate long-term effects[115] and apply standardized reporting methods such as TRIPOD-AI[134], which is currently under development.

Another important issue to consider is the quality measure that is applied when evaluating a CPATH algorithm—in other words, when is an algorithm good enough? Typically, studies in which CPATH algorithms aim to produce a diagnosis comparable to those used by a pathologist will compare the CPATH algorithm with scores from a panel of pathologists, often concluding that the algorithm may be applied in clinical practice if the performance is close to the average of the pathologists. However, as argued by Campanella and colleagues[75], a clinically useful decision support system should ideally take into account the fact that, in a real-world setting, pathologists do not evaluate images as an isolated task but rather can opt to use IHC and consultation with colleagues as part of their diagnostic workup, if deemed necessary. Campanella and colleagues conclude that "achieving 100% sensitivity with an acceptable false positive rate" should be the aim to achieve clinical-grade CPATH algorithms. Rather than defining a threshold for clinical usefulness, conclusions about the true clinical value of CPATH can be drawn only from prospective trials, which incorporate the entire diagnostic process, including the use of existing reporting standards[115,131]. Before such usefulness can be demonstrated, far-ranging conclusions about the impact of CPATH on diagnostics should be avoided, as they may lead to an overly optimistic view.

Future directions

Even though promising results for deep learning CPATH algorithms have been shown in many studies, it is still too early to distinguish the hope from the hype. Although the hope is motivated by the development of CPATH algorithms of increasing accuracy across many fields in pathology that have the potential to help pathologists in their clinical practice, the hype often leads to the question of whether AI will replace pathologists. Given this question, it is important to realize the breadth of tasks that a pathologist performs: pathologists do not simply analyze a piece of tissue under a microscope; they also integrate information from different sources of clinical data, their own understanding of the disease, the diagnostic process and the specific circumstances of the patient, and then communicate and explain the outcomes of the analysis for both other clinicians and, increasingly, for patients. Thus, it is important to stress that pathologists are not likely to be replaced with AI algorithms anytime soon. What could be achieved relatively soon is AI algorithms that work in conjunction with pathologists, rather than as stand-alone solutions, to remove the need for tedious, repetitive work, such as identifying lymph node metastases[5], or to increase the quality of diagnostic grading[4,66]. In this context, it is important to differentiate between most high-income countries, such as the USA or the Netherlands, and low- or middle-income countries, such as China or India. The former generally have sufficient pathologists for their current workload (although the issue of workloads is expected to become problematic in the near future), whereas access to pathological expertise is challenging and sometimes even impossible in the latter. In the absence of pathologists, algorithms could yield urgently needed data to inform diagnoses, which would be an important step forward. Obviously, the infrastructure around digital pathology in some

settings, such as rural hospitals in low- and middle-income countries, would present challenges, and it is worth noting recent initiatives to address this limitation by giving access to CPATH algorithms without the need for a full digital pathology infrastructure[135].

Explainable AI. CPATH solutions based on deep learning models are often described as black boxes, indicating that, because of the nature of these systems (being trained rather than explicitly programmed), it is very difficult for humans to understand the exact underlying functioning of the system[113]. As a result, correcting certain erroneous behaviors may be more difficult, and acceptance by humans (as well as regulatory approval) may be hampered[114]. This problem has given rise to research on explainable AI, in which techniques are developed that enable better understanding of the functioning of deep learning models. The current state of the art of methods that can shed light inside the black box has been extensively reviewed elsewhere[136]. Interestingly, the authors of that survey[136] conclude that there is no consensus on the exact meaning of the term explainable, as it has different requirements in different contexts and for different stakeholders. Although techniques for improving the explainability of AI will support acceptance by the community, the emphasis that should be placed on precise understanding of the mechanics of the technology, rather than on the functioning of the system as a whole in the context in which it is used[137], is debatable, especially when such systems are integrated into the clinical workflow of pathology diagnostics. Rigorous validation and quality assurance and checking procedures will be critical to prove the correct functioning of CPATH solutions, both at initial market entrance and also after future updates. This topic is highly relevant to regulatory bodies such as the Food and Drug Administration, which has recently proposed a regulatory framework[138] to implement a predetermined change control plan, in which manufacturers have to explain what aspects they intend to change through learning, and how the algorithm will learn and change while remaining safe and effective, as well as strategies to mitigate performance loss.

Ethics. The use of patient data and the deployment of machines that aid diagnostics, potentially even in a (partly) autonomous fashion, lead to a number of ethical concerns. The development of CPATH solutions requires large amounts of data (both images and associated metadata). The use of human data for health-care research and product development in general creates ethical and legal challenges that have to be addressed properly. Respecting patient privacy and obtaining approval for use of data are important requirements to comply with. Unfortunately, from a practical standpoint, these requirements may reduce the options for the reuse of existing data for AI development, and may lead to increased costs to arrive at the required numbers of cases. A careful balancing between privacy protection and the benefits of data-driven innovation is needed[139] that requires the involvement of all stakeholders[140]. In addition, the collection of data at the scale needed for CPATH development (thousands of cases), which can be made publicly available, complicates matters even further. Aside from the danger of data breaches, collecting large amounts of data may enable researchers to make connections that were previously not possible, potentially putting patient privacy at risk even if data are collected in an anonymized manner[114,139]. An alternative to establishing large, multicenter datasets in a central location for machine learning is the application of so-called federated learning strategies. With federated learning, the procedure for training the machine learning models is adapted so that it can deal with data residing in separate locations, obviating the need to bring the data together and thereby circumventing some of the problems described above[141,142].

In 2018, a special expert group of the European Commission published a set of ethics guidelines for trustworthy AI[143] that detailed a framework to help achieve AI solutions that are lawful, ethical and robust. An important conclusion from the guideline is that "trustworthy AI is not about ticking boxes, but about continuously identifying and implementing requirements, evaluating solutions, ensuring improved outcomes throughout the AI system's life cycle, and involving stakeholders in this"[143]. How the establishment of data collections may result in biases that, if used for AI development, can amplify injustices in society has been described extensively[115,117]. Such algorithmic bias is not directly the consequence of AI model development, but the wide-scale deployment of such models may compound "existing inequities in socioeconomic status, race, ethnic background, religion, gender, disability or sexual orientation"[117]. As it is not possible to recognize inequities in data collection a priori, the European Commission's recommendation to continuously engage in discussions with all stakeholders is even more imperative[143].

Conclusion

Promising results of the application of AI to histopathological images have provoked a large number of research studies, now resulting in CPATH solutions that have comparable performance to pathologists for several specific diagnostic tasks. In addition to the use of AI to conduct human expert diagnostic tasks, we have only begun to scratch the surface with respect to the use of AI for discovery of prognostic features, prediction of therapy success or assessment of the relation between the morphological phenotypes of disease and genotypes. Whereas many technical challenges have been overcome, clinical usefulness has not been proved yet and several hurdles still have to be overcome. Next to the challenge of collecting sufficiently large sets of annotated WSIs, prospective studies have to be conducted to show the true benefit of AI for histopathological diagnostics. Issues related to explainability, ethics and regulation are also insufficiently studied and will require more attention in the near future. Even though the field is not fully matured yet, we expect CPATH to play a dominant role in the future of histopathology, making diagnostics more efficient and accurate, helping pathologists meet the requirements of an increasing number of patients and the need for more extensive and accurate histopathological assessment to aid the increasing spectrum of treatment options for many diseases.

Whole Slide Image (WSI)

Received: 18 February 2020; Accepted: 31 March 2021;
Published online: 14 May 2021

References

1. Yu, K. H., Beam, A. L. & Kohane, I. S. Artificial intelligence in healthcare. *Nat. Biomed. Eng.* **2**, 719–731 (2018).
2. Hamet, P. & Tremblay, J. Artificial intelligence in medicine. *Metabolism* **69**, S36–S40 (2017).
3. Litjens, G. et al. A survey on deep learning in medical image analysis. *Med. Image Anal.* **42**, 60–88 (2017).
4. Bulten, W. et al. Automated deep-learning system for Gleason grading of prostate cancer using biopsies: a diagnostic study. *Lancet Oncol.* **21**, 233–241 (2020).
5. Ehteshami Bejnordi, B. et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA* **318**, 2199–2210 (2017).
6. Bera, K., Schalper, K. A., Rimm, D. L., Velcheti, V. & Madabhushi, A. Artificial intelligence in digital pathology – new tools for diagnosis and precision oncology. *Nat. Rev. Clin. Oncol.* **16**, 703–715 (2019).
7. Fuchs, T. J. & Buhmann, J. M. Computational pathology: challenges and promises for tissue analysis. *Comput. Med. Imaging Graph.* **35**, 515–530 (2011).
8. Louis, D. N. et al. Computational pathology: an emerging definition. *Arch. Pathol. Lab. Med.* **138**, 1133–1138 (2014).
9. Mendelsohn, M. L., Kolman, W. A., Perry, B. & Prewitt, J. M. Computer analysis of cell images. *Postgrad. Med.* **38**, 567–573 (1965).
10. Zwanenburg, A. et al. The image biomarker standardization initiative: standardized quantitative radiomics for high-throughput image-based phenotyping. *Radiology* **295**, 328–338 (2020).
11. Beck, A. H. et al. Systematic analysis of breast cancer morphology uncovers stromal features associated with survival. *Sci. Transl. Med.* **3**, 108–113 (2011).

12. Kather, J. N. et al. Multi-class texture analysis in colorectal cancer histology. *Sci. Rep.* **6**, 27988 (2016).
13. Madabhushi, A. & Lee, G. Image analysis and machine learning in digital pathology: challenges and opportunities. *Med. Image Anal.* **33**, 170–175 (2016).
14. Srinidhi, C. L., Ciga, O. & Martel, A. L. Deep neural network models for computational histopathology: a survey. *Med. Image Anal.* **67**, 101813 (2021).
15. Cireşan, D. C., Meier, U., Masci, J., Gambardella, L. M. & Schmidhuber, J. Flexible, high performance convolutional neural networks for image classification. In *Proc. 22nd International Joint Conference on Artificial Intelligence* 1237–1242 (2011).
16. Cireşan, D. C., Giusti, A., Gambardella, L. M. & Schmidhuber, J. Mitosis detection in breast cancer histology images with deep neural networks. In *Proc. Medical Image Computing and Computer-Assisted Intervention, Lecture Notes in Computer Science* Vol. 8150, 411–418 (2013).
17. Cruz-Roa, A. et al. Automatic detection of invasive ductal carcinoma in whole slide images with convolutional neural networks. In *Proc. SPIE Medical Imaging* Vol. 9041, 904103 (2014).
18. Ertosun, M. G. & Rubin, D. L. Automated grading of gliomas using deep learning in digital pathology images: a modular approach with ensemble of convolutional neural networks. In *Proc. American Medical Informatics Association Annual Symposium* 1899–1908 (2015).
19. Wong, G. L. et al. Artificial intelligence in prediction of non-alcoholic fatty liver disease and fibrosis. *J. Gastroenterol. Hepatol.* **36**, 543–550 (2021).
20. Hermsen, M. et al. Deep learning–based histopathologic assessment of kidney tissue. *J. Am. Soc. Nephrol.* **30**, 1968–1979 (2019).
21. Litjens, G. et al. Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. *Sci. Rep.* **6**, 26286 (2016).
22. Litjens, G. et al. 1399 H&E-stained sentinel lymph node sections of breast cancer patients: the CAMELYON dataset. *GigaScience* **7**, giy065 (2018).
23. Liu, Y. et al. Detecting cancer metastases on gigapixel pathology images. Preprint at https://arxiv.org/abs/1703.02442 (2017).
24. White House Office of Science and Technology Policy. Preparing for the Future of Artificial Intelligence (2016); https://obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/preparing_for_the_future_of_ai.pdf
25. Wang, D., Khosla, A., Gargeya, R., Irshad, H. & Beck, A. H. Deep learning for identifying metastatic breast cancer. Preprint at https://arxiv.org/abs/1606.05718 (2016).
26. Wang, X. et al. Weakly supervised deep learning for whole slide lung cancer image analysis. *IEEE Trans. Cybern.* **50**, 3950–3962 (2019).
27. Syrykh, C. et al. Accurate diagnosis of lymphoma on whole-slide histopathology images using deep learning. *NPJ Digital Med.* **3**, 63 (2020).
28. Tabibu, S., Vinod, P. K. & Jawahar, C. V. Pan-renal cell carcinoma classification and survival prediction from histopathology images using deep learning. *Sci. Rep.* **9**, 10509 (2019).
29. Sari, C. T. & Gunduz-Demir, C. Unsupervised feature extraction via deep learning for histopathological classification of colon tissue images. *IEEE Trans. Med. Imaging* **38**, 1139–1149 (2019).
30. Rawat, R. R. et al. Deep learned tissue 'fingerprints' classify breast cancers by ER/PR/Her2 status from H&E images. *Sci. Rep.* **10**, 7275 (2020).
31. Lee, B. & Paeng, K. A robust and effective approach towards accurate metastasis detection and pN-stage classification in breast cancer. In *Proc. Medical Image Computing and Computer Assisted Intervention, Lecture Notes in Computer Science* Vol. 11071, 841–850 (2018).
32. Awan, R., Koohbanani, N. A., Shaban, M., Lisowska, A. & Rajpoot, N. Context-aware learning using transferable features for classification of breast cancer histology images. In *Proc. International Conference on Image Analysis and Recognition* 788–795 (2018).
33. Feng, Y., Zhang, L. & Mo, J. Deep manifold preserving autoencoder for classifying breast cancer histopathological images. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **17**, 91–101 (2020).
34. Galateau Salle, F. et al. Comprehensive molecular and pathologic evaluation of transitional mesothelioma assisted by deep learning approach: a multi-institutional study of the international mesothelioma panel from the MESOPATH Reference Center. *J. Thorac. Oncol.* **15**, 1037–1053 (2020).
35. Iizuka, O. et al. Deep learning models for histopathological classification of gastric and colonic epithelial tumours. *Sci. Rep.* **10**, 1504 (2020).
36. Kiani, A. et al. Impact of a deep learning assistant on the histopathologic classification of liver cancer. *NPJ Digital Med.* **3**, 23 (2020).
37. Kwok, S. Multiclass classification of breast cancer in whole-slide images. In *Proc. International Conference on Image Analysis and Recognition* 931–940 (2018).
38. Wei, J. W. et al. Pathologist-level classification of histologic patterns on resected lung adenocarcinoma slides with deep neural networks. *Sci. Rep.* **9**, 3358 (2019).
39. Yang, H., Kim, J. Y., Kim, H. & Adhikari, S. P. Guided soft attention network for classification of breast cancer histopathology images. *IEEE Trans. Med. Imaging* **39**, 1306–1315 (2020).
40. Pinckaers, H. & Litjens, G. Neural ordinary differential equations for semantic segmentation of individual colon glands. Preprint at https://arxiv.org/abs/1910.10470 (2019).
41. Naylor, P., Lae, M., Reyal, F. & Walter, T. Segmentation of nuclei in histopathology images by deep regression of the distance map. *IEEE Trans. Med. Imaging* **38**, 448–459 (2019).
42. Long, F. Microscopy cell nuclei segmentation with enhanced U-Net. *BMC Bioinformatics* **21**, 8 (2020).
43. Jia, Z., Huang, X., Chang, E. I.-C. & Xu, Y. Constrained deep weak supervision for histopathology image segmentation. *IEEE Trans. Med. Imaging* **36**, 2376–2388 (2017).
44. Graham, S. et al. MILD-Net: minimal information loss dilated network for gland instance segmentation in colon histology images. *Med. Image Anal.* **52**, 199–211 (2019).
45. Graham, S. et al. Hover-Net: simultaneous segmentation and classification of nuclei in multi-tissue histology images. *Med. Image Anal.* **58**, 101563 (2019).
46. Agarwalla, A., Shaban, M. & Rajpoot, N. M. Representation-aggregation networks for segmentation of multi-gigapixel histology images. Preprint at https://arxiv.org/abs/1707.08814 (2017).
47. Bueno, G., Fernandez-Carrobles, M. M., Gonzalez-Lopez, L. & Deniz, O. Glomerulosclerosis identification in whole slide images using semantic segmentation. *Comput. Methods Prog. Biomed.* **184**, 105273 (2020).
48. Chen, H. et al. DCAN: deep contour-aware networks for object instance segmentation from histology images. *Med. Image Anal.* **36**, 135–146 (2017).
49. de Bel, T. et al. Automatic segmentation of histopathological slides of renal tissue using deep learning. in *Proc. SPIE Medical Imaging Digital Pathology*, 1058112 (2018); https://doi.org/10.1117/12.2293717
50. Xu, G. et al. CAMEL: a weakly supervised learning framework for histopathology image segmentation. In *Proc. International Conference on Computer Vision* 10681–10690 (2019).
51. Sirinukunwattana, K. et al. Locality sensitive deep learning for detection and classification of nuclei in routine colon cancer histology images. *IEEE Trans. Med. Imaging* **35**, 1196–1206 (2016).
52. Swiderska-Chadaj, Z. et al. Learning to detect lymphocytes in immunohistochemistry with deep learning. *Med. Image Anal.* **58**, 101547 (2019).
53. Le, H. et al. Utilizing automated breast cancer detection to identify spatial distributions of tumor-infiltrating lymphocytes in invasive breast cancer. *Am. J. Pathol.* **190**, 1491–1504 (2020).
54. Akbar, S. et al. Automated and manual quantification of tumour cellularity in digital slides for tumour burden assessment. *Sci. Rep.* **9**, 14099 (2019).
55. Hou, L. et al. Sparse autoencoder for unsupervised nucleus detection and representation in histopathology images. *Pattern Recognit.* **86**, 188–200 (2019).
56. Veta, M. et al. Predicting breast tumor proliferation from whole-slide images: the TUPAC16 challenge. *Med. Image Anal.* **54**, 111–121 (2019).
57. Tellez, D. et al. Whole-slide mitosis detection in H&E breast histology using PHH3 as a reference to train distilled stain-invariant convolutional networks. *IEEE Trans. Med. Imaging* **37**, 2126–2136 (2018).
58. Mahmood, T., Arsalan, M., Owais, M., Lee, M. B. & Park, K. R. Artificial intelligence-based mitosis detection in breast cancer histopathology images using faster R-CNN and deep CNNs. *J. Clin. Med.* **9**, 749 (2020).
59. Chen, H., Wang, X. & Heng, P. A. Automated mitosis detection with deep regression networks. In *Proc. IEEE International Symposium on Biomedical Imaging* 1204–1207 (2016).
60. Li, C. et al. Weakly supervised mitosis detection in breast histopathology images using concentric loss. *Med. Image Anal.* **53**, 165–178 (2019).
61. Nagpal, K. et al. Development and validation of a deep learning algorithm for improving Gleason scoring of prostate cancer. *NPJ Digital Med.* **2**, 48 (2019).
62. Jansen, I. et al. Automated detection and grading of non-muscle-invasive urothelial cell carcinoma of the bladder. *Am. J. Pathol.* **190**, 1483–1490 (2020).
63. Karimi, D. et al. Deep learning-based Gleason grading of prostate cancer from histopathology images—role of multiscale decision aggregation and data augmentation. *IEEE J. Biomed. Health Inform.* **24**, 1413–1426 (2020).
64. Korde, L. A. et al. Neoadjuvant chemotherapy, endocrine therapy, and targeted therapy for breast cancer: ASCO guideline. *J. Clin. Oncol.* https://doi.org/10.1200/JCO.20.03399 (2021).
65. Balkenhol, M. et al. Deep learning assisted mitotic counting for breast cancer. *Lab. Invest.* **99**, 1596–1606 (2019).
66. Ström, P. et al. Artificial intelligence for diagnosis and grading of prostate cancer in biopsies: a population-based, diagnostic study. *Lancet Oncol.* **21**, 222–232 (2020).
67. Gertych, A. et al. Convolutional neural networks can accurately distinguish four histologic growth patterns of lung adenocarcinoma in digital slides. *Sci. Rep.* **9**, 1483 (2019).
68. Salgado, R. et al. The evaluation of tumor-infiltrating lymphocytes (TILs) in breast cancer: recommendations by an International TILs Working Group 2014. *Ann. Oncol.* **26**, 259–271 (2015).

69. Saltz, J. et al. Spatial organization and molecular correlation of tumor-infiltrating lymphocytes using deep learning on pathology images. *Cell Rep.* **23**, 181–193 (2018).
70. AbdulJabbar, K. Geospatial immune variability illuminates differential evolution of lung adenocarcinoma. *Nat. Med.* **26**, 1054–1062 (2020).
71. Kather, J. N. et al. Predicting survival from colorectal cancer histology slides using deep learning: a retrospective multicenter study. *PLoS Med.* **16**, e1002730 (2019).
72. Geessink, O. G. F. et al. Computer aided quantification of intratumoral stroma yields an independent prognosticator in rectal cancer. *Cell. Oncol.* **42**, 331–341 (2019).
73. Kapil, A. et al. DASGAN–joint domain adaptation and segmentation for the analysis of epithelial regions in histopathology PD-L1 images. Preprint at https://arxiv.org/abs/1906.11118 (2019).
74. Sha, L. et al. Multi-field-of-view deep learning model predicts non small cell lung cancer programmed death-ligand 1 status from whole-slide hematoxylin and eosin images. *J. Pathol. Inform.* **10**, 24 (2019).
75. Campanella, G. et al. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nat. Med.* **25**, 1301–1309 (2019).
76. Zhang, Z. et al. Pathologist-level interpretable whole-slide cancer diagnosis with deep learning. *Nat. Mach. Intell.* **1**, 236–245 (2019).
77. Zhou, C. et al. Histopathology classification and localization of colorectal cancer using global labels by weakly supervised deep learning. *Comput. Med. Imaging Graph.* **88**, 101861 (2021).
78. Song, Z. et al. Clinically applicable histopathological diagnosis system for gastric cancer detection using deep learning. *Nat. Commun.* **11**, 4294 (2020).
79. Albarqouni, S. et al. AggNet: deep learning from crowds for mitosis detection in breast cancer histology images. *IEEE Trans. Med. Imaging* **35**, 1313–1321 (2016).
80. Amgad, M. et al. Structured crowdsourcing enables convolutional segmentation of histology images. *Bioinformatics* **35**, 3461–3467 (2019).
81. Bulten, W. et al. Epithelium segmentation using deep learning in H&E-stained prostate specimens with immunohistochemistry as reference standard. *Sci. Rep.* **9**, 864 (2019).
82. Valkonen, M. et al. Cytokeratin-supervised deep learning for automatic recognition of epithelial cells in breast cancers stained for ER, PR, and Ki-67. *IEEE Trans. Med. Imaging* **39**, 534–542 (2020).
83. Alemi Koohbanani, N., Jahanifar, M., Zamani Tajadin, N. & Rajpoot, N. NuClick: a deep learning framework for interactive segmentation of microscopic images. *Med. Image Anal.* **65**, 101771 (2020).
84. Bokhorst, J. M. et al. Learning from sparsely annotated data for semantic segmentation in histopathology images. In *Proc. International Conference on Medical Imaging with Deep Learning, Proceedings of Machine Learning Research* Vol. 102, 84–91 (2019).
85. Brieu, N. et al. Domain adaptation-based augmentation for weakly supervised nuclei detection. Preprint at https://arxiv.org/abs/1907.04681 (2019).
86. Gadermayr, M., Gupta, L., Klinkhammer, B. M., Boor, P. & Merhof, D. Unsupervisedly training GANs for segmenting digital pathology with automatically generated annotations. In *Proc. International Conference on Medical Imaging with Deep Learning, Proceedings of Machine Learning Research* Vol. 102, 175–184 (2019).
87. Liang, Q. et al. Weakly supervised biomedical image segmentation by reiterative learning. *IEEE J. Biomed. Health Inform.* **23**, 1205–1214 (2019).
88. Gao, S. et al. Using case-level context to classify cancer pathology reports. *PLoS ONE* **15**, e0232840 (2020).
89. Alawad, M. et al. Automatic extraction of cancer registry reportable information from free-text pathology reports using multitask convolutional neural networks. *J. Am. Med. Inform. Assoc.* **27**, 89–98 (2020).
90. Coudray, N. et al. Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. *Nat. Med.* **24**, 1559–1567 (2018).
91. Pawlowski, N. et al. Needles in haystacks: on classifying tiny objects in large images. Preprint at https://arxiv.org/abs/1908.06037 (2019).
92. Ilse, M., Tomczak, J. M. & Welling, M. Attention-based deep multiple instance learning. In *Proc. International Conference on Machine Learning, Proceedings of Machine Learning Research* Vol. 80, 2127–2136 (2018).
93. Hou, L. et al. Patch-based convolutional neural network for whole slide tissue image classification. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* 2424–2433 (2016).
94. Lu, M. Y. et al. Data efficient and weakly supervised computational pathology on whole slide images. *Nat. Biomed. Eng.* https://doi.org/10.1038/s41551-020-00682-w (2021).
95. Tellez, D., Litjens, G., van der Laak, J. & Ciompi, F. Neural image compression for gigapixel histopathology image analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* **43**, 567–578 (2021).
96. Pinckaers, J. H. F. M., van Ginneken, B. & Litjens, G. Streaming convolutional neural networks for end-to-end learning with multi-megapixel images. *IEEE Trans. Pattern Anal. Mach. Intell.* https://doi.org/10.1109/TPAMI.2020.3019563 (2020).
97. Wulczyn, E. et al. Deep learning-based survival prediction for multiple cancer types using histopathology images. *PLoS ONE* **15**, e0233678 (2020).
98. Skrede, O. J. Deep learning for prediction of colorectal cancer outcome: a discovery and validation study. *Lancet* **395**, 350–360 (2020).
99. Saillard, C. et al. Predicting survival after hepatocellular carcinoma resection using deep-learning on histological slides. *Hepatology* **72**, 2000–2013 (2020).
100. Qaiser, T. et al. Digital tumor-collagen proximity signature predicts survival in diffuse large B-cell lymphoma. In *Proc. European Congress on Digital Pathology, Lecture Notes in Computer Science* Vol. 11435, 163–171 (2019).
101. Mobadersany, P. et al. Predicting cancer outcomes from histology and genomics using convolutional networks. *Proc. Natl Acad. Sci. USA* **115**, E2970–E2979 (2018).
102. Bychkov, D. et al. Deep learning based tissue analysis predicts outcome in colorectal cancer. *Sci. Rep.* **8**, 3395 (2018).
103. Courtiol, P. et al. Deep learning-based classification of mesothelioma improves prediction of patient outcome. *Nat. Med.* **25**, 1519–1525 (2019).
104. Kulkarni, P. M. et al. Deep learning based on standard H&E images of primary melanoma tumors identifies patients at risk for visceral recurrence and death. *Clin. Cancer Res.* **26**, 1126–1134 (2020).
105. Cui, D., Liu, Y., Liu, G. & Liu, L. A multiple-instance learning-based convolutional neural network model to detect the *IDH1* mutation in the histopathology images of glioma tissues. *J. Comput. Biol.* **27**, 1264–1272 (2020).
106. Liu, S. et al. Isocitrate dehydrogenase (IDH) status prediction in histopathology images of gliomas using deep learning. *Sci. Rep.* **10**, 7733 (2020).
107. Kather, J. N. et al. Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. *Nat. Med.* **25**, 1054–1056 (2019).
108. Kather, J. N. et al. Pan-cancer image-based detection of clinically actionable genetic alterations. *Nat. Cancer* **1**, 789–799 (2020).
109. Fu, Y. et al. Pan-cancer computational histopathology reveals mutations, tumor composition and prognosis. *Nat. Cancer* **1**, 800–810 (2020).
110. Schmauch, B. et al. A deep learning model to predict RNA-Seq expression of tumours from whole slide images. *Nat. Commun.* **11**, 3877 (2020).
111. Couture, H. D. et al. Image analysis with deep learning to predict breast cancer grade, ER status, histologic subtype, and intrinsic subtype. *NPJ Breast Cancer* **4**, 30 (2018).
112. Sirinukunwattana, K. et al. Image-based consensus molecular subtype (imCMS) classification of colorectal cancer using deep learning. *Gut* **70**, 544–554 (2021).
113. Durán, J. M. & Jongsma, K. R. Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI. *J. Med Ethics* https://doi.org/10.1136/medethics-2020-106820 (2021).
114. Abels, E. et al. Computational pathology definitions, best practices, and recommendations for regulatory guidance: a white paper from the Digital Pathology Association. *J. Pathol.* **249**, 286–294 (2019).
115. Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G. & King, D. Key challenges for delivering clinical impact with artificial intelligence. *BMC Med.* **17**, 195 (2019).
116. McKinney, S. M. et al. International evaluation of an AI system for breast cancer screening. *Nature* **577**, 89–94 (2020).
117. Panch, T., Mattie, H. & Atun, R. Artificial intelligence and algorithmic bias: implications for health systems. *J. Glob. Health* **9**, 010318 (2019).
118. de Bel, T., Hermsen, M., Kers, J., van der Laak, J. & Litjens, G. J. S. Stain-transforming cycle-consistent generative adversarial networks for improved segmentation of renal histopathology. In *Proc. International Conference on Medical Imaging with Deep Learning, Proceedings of Machine Learning Research* Vol. 102, 151–163 (2019).
119. Liu, Y. et al. Artificial intelligence-based breast cancer nodal metastasis detection: insights into the black box for pathologists. *Arch. Pathol. Lab. Med.* **143**, 859–868 (2019).
120. Tellez, D. et al. Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology. *Med. Image Anal.* **58**, 101544 (2019).
121. Cho, H., Lim, S., Choi, G. & Min, H. Neural stain-style transfer learning using GAN for histopathological images. Preprint at https://arxiv.org/abs/1710.08543 (2017).
122. Janowczyk, A., Basavanhally, A. & Madabhushi, A. Stain Normalization using Sparse AutoEncoders (StaNoSA): application to digital pathology. *Comput. Med. Imaging Graph.* **57**, 50–61 (2017).

123. Shaban, M. T., Baur, C., Navab, N. & Albarqouni, S. StainGAN: stain style transfer for digital histological images. In *Proc. IEEE International Symposium on Biomedical Imaging* 953–956 (2019).
124. Zheng, Y. et al. Stain standardization capsule for application-driven histopathological image normalization. *IEEE J. Biomed. Health Inform.* **25**, 337–347 (2021).
125. Linmans, J., van der Laak, J. & Litjens, G. Efficient out-of-distribution detection in digital pathology using multi-head convolutional neural networks. In *Proc. Conference on Medical Imaging with Deep Learning, Proceedings of Machine Learning Research* Vol. 121, 465–478 (2020).
126. Kohl, S. et al. A probabilistic U-Net for segmentation of ambiguous images. *Adv. Neural Inf. Process. Syst.* (2018).
127. Kleppe, A. et al. Designing deep learning studies in cancer diagnostics. *Nat. Rev. Cancer* **21**, 199–211 (2021).
128. Staartjes, V. E. & Kernbach, J. M. Significance of external validation in clinical machine learning: let loose too early. *Spine J.* **20**, 1159–1160 (2020).
129. Beede, E. et al. A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy. In *Proc. CHI Conference on Human Factors in Computing Systems* 1–12 (2020).
130. Dudgeon, S. N. et al. A pathologist-annotated dataset for validating artificial intelligence: a project description and pilot study. Preprint at https://arxiv.org/abs/2010.06995 (2020).
131. Nagendran, M. et al. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. *BMJ* **368**, m689 (2020).
132. Allen, T. C. Regulating artificial intelligence for a successful pathology future. *Arch. Pathol. Lab. Med.* **143**, 1175–1179 (2019).
133. Dong, J. et al. Clinical trials for artificial intelligence in cancer diagnosis: a cross-sectional study of registered trials in ClinicalTrials.gov. *Front. Oncol.* **15**, 1629 (2020).
134. Collins, G. S. & Moons, K. G. M. Reporting of artificial intelligence prediction models. *Lancet* **393**, 1577–1579 (2019).
135. Chen, P. H. C. et al. An augmented reality microscope with real-time artificial intelligence integration for cancer diagnosis. *Nat. Med.* **25**, 1453–1457 (2019).
136. Guidotti, R. et al. A survey of methods for explaining black box models. *ACM Comput. Surv.* **51**, 93 (2019).
137. Kroll, J. A. The fallacy of inscrutability. *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.* **376**, 20180084 (2018).
138. US Food and Drug Administration (FDA). Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD)—Discussion Paper and Request for Feedback. https://www.fda.gov/files/medicaldevices/published/US-FDA-Artificial-Intelligence-and-Machine-Learning-Discussion-Paper.pdf (accessed 3 May, 2021).
139. Price, W. N. & Cohen, I. G. Privacy in the age of medical big data. *Nat. Med.* **25**, 37–43 (2019).
140. Laï, M. C., Brian, M. & Mamzer, M. F. Perceptions of artificial intelligence in healthcare: findings from a qualitative survey study among actors in France. *J. Transl. Med.* **18**, 14 (2020).
141. Rieke, N. et al. The future of digital health with federated learning. *NPJ Digit. Med.* **3**, 119 (2020).
142. Sheller, M. J. et al. Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. *Sci. Rep.* **10**, 12598 (2020).
143. European Commission. *Ethics Guidelines for Trustworthy AI* (2019); https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai
144. Sirinukunwattana, K. et al. Gland segmentation in colon histology images: the glas challenge contest. *Med. Image Anal.* **35**, 489–502 (2017).
145. Ronneberger, O., Fischer, P. & Brox, T. U-Net: convolutional networks for biomedical image segmentation. In *Proc. Medical Image Computing and Computer-Assisted Intervention, Lecture Notes in Computer Science* Vol. 9351, 234–241 (2015).

Acknowledgements

We thank M. Hermsen for providing Fig. 1. J.v.d.L. acknowledges funding from the Knut and Alice Wallenberg Foundation, Sweden, and received funding from the Innovative Medicines Initiative 2 Joint Undertaking under grant agreement no. 945358. This Joint Undertaking receives support from the European Union's Horizon 2020 research and innovation program and EFPIA. G.L. acknowledges funding from the Dutch Cancer Society (KUN 2015-7970) and the Netherlands Organization for Scientific Research (NWO; project number 016.186.152). F.C. acknowledges funding from the European Union's Horizon 2020 research and innovation program under grant agreement no. 825292 (ExaMode, http://www.examode.eu/); the Dutch Cancer Society (KWF; project no. 11917); and the Netherlands Organization for Scientific Research (NWO; project no. 18388).

Author contributions

All authors were involved in identifying relevant literature, and in drafting and revising the manuscript.

Competing interests

J.v.d.L. is a member of the advisory boards of Philips, the Netherlands, and ContextVision, Sweden, and received research funding from Philips, the Netherlands, ContextVision, Sweden, and Sectra, Sweden, in the last 5 years. G.L. is a member of the Medical Advances Advisory Board of Vital Images (Minnetonka, USA), received research funding from Philips Digital Pathology Solutions (Best, the Netherlands) and had a consultancy role for Novartis (Basel, Switzerland). F.C. is chair of the Scientific and Medical Advisory Board of TRIBVN Healthcare (Paris, France).

Additional information

Correspondence should be addressed to J.v.d.L.

Peer review information *Nature Medicine* thanks Richard Levenson, Nasir Rajpoot and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Joao Monteiro was the primary editor on this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.