JIM Review

doi: 10.1111/joim.13030

Artificial intelligence as the next step towards precision pathology

B. Acs[1], M. Rantalainen[2] & J. Hartman[1]

From the [1]Department of Oncology and Pathology; and [2]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden

Abstract. Acs B, Rantalainen M, Hartman J (Karolinska Institutet, Stockholm, Sweden) Artificial intelligence as the next step towards precision pathology. *J Intern Med*; 2020; **288**: 62–81.

Pathology is the cornerstone of cancer care. The need for accuracy in histopathologic diagnosis of cancer is increasing as personalized cancer therapy requires accurate biomarker assessment. The appearance of digital image analysis holds promise to improve both the volume and precision of histomorphological evaluation. Recently, machine learning, and particularly deep learning, has enabled rapid advances in computational pathology. The integration of machine learning into routine care will be a milestone for the healthcare sector in the next decade, and histopathology is right at the centre of this revolution. Examples of potential high-value machine learning applications include both model-based assessment of routine diagnostic features in pathology, and the ability to extract and identify novel features that provide insights into a disease. Recent groundbreaking results have demonstrated that applications of machine learning methods in pathology significantly improves metastases detection in lymph nodes, Ki67 scoring in breast cancer, Gleason grading in prostate cancer and tumour-infiltrating lymphocyte (TIL) scoring in melanoma. Furthermore, deep learning models have also been demonstrated to be able to predict status of some molecular markers in lung, prostate, gastric and colorectal cancer based on standard HE slides. Moreover, prognostic (survival outcomes) deep neural network models based on digitized HE slides have been demonstrated in several diseases, including lung cancer, melanoma and glioma. In this review, we aim to present and summarize the latest developments in digital image analysis and in the application of artificial intelligence in diagnostic pathology.

Keywords: artificial intelligence, deep learning, digital image analysis, digital pathology, machine learning, pathology.

Introduction

The importance of anatomic pathology to diagnose and classify disease cannot be underestimated. The pathologist's diagnosis on histological slides is at the centre of diagnosis, for clinical and pharmaceutical research and, most importantly, for decision-making on how to treat cancer patients in the daily practice. The need for accuracy in histopathologic diagnosis of cancer is increasing as personalized therapy requires accurate biomarker assessment [1]. However, most of the world is facing with an urgent shortage of pathologists [2]. Furthermore, despite the fact that robust guidelines for optimization are in use since several years [3, 4], biomarker assessment is still limited by the subjectivity linked to (i) defining the appropriate tumour areas to investigate within a heterogenous tissue section; and (ii) the visual interpretation of biomarker distribution and intensity patterns in tumour cells and stromal tissues [5]. Several studies have shown low between-laboratory, inter-observer and intra-observer reproducibility in biomarker evaluation [6, 7]. This variability is hindering both the process of discovering new biomarkers and their utilization in clinical practice. Computational image analysis in pathology has been around for many years [8, 9]. However, its application in routine pathology has been confined, due to the limited capacities of glass slide digitization, computer hardware, processing time and image analysis methods as well as data storage. Besides, qualitative or semi-quantitative manual visual assessment has been considered adequate, as therapeutic decision-making was not defined to rely on quantitative diagnostic results in many

cases. Recently, rapid development of digital microscopy has enabled digitalization of histological slides at high-resolution and high speed, which can now firmly support training, research and diagnostics in pathology. The appearance of digital image analysis (DIA) algorithms holds promise to improve the volume precision of histomorphological evaluation. Moreover, digital pathology has recently received increasing attention, partially due to the competition emerging from molecular profiling platforms that deliver precise quantitative results with minimal inter-observer variability problems. This has increased the demand for routine histopathological assessment to keep up with the high-throughput precision diagnostics. Therefore, systematic computational pathology research initiatives have been presented to accelerate the quantitative assessment of both morphological patterns and biomarker expression in histopathology [10]. The most promising and fundamental advances in computational pathology is based on artificial intelligence (AI) and machine learning methodologies, which delivers computer models with image recognition that match, or outperform, human experts. First, the terms artificial intelligence, machine learning and deep learning should be clarified as they have often been used interchangeably. AI is an umbrella term enclosing the methods for a computer to emulate, or exceed in some extent, human intelligence [11]. Machine learning is a subfield of AI that applies statistical methods to optimize models for a specific task without depending on specific human directions to define all of the rules or parameters in the model. Currently, supervised learning is dominating AI and ML applications in the medical domain. Supervised learning is based on the principle of optimizing (i.e. 'training') a model using training data (e.g. medical images) that have labels available (e.g. clinical classification, patient outcomes or pixel-level image annotations). Deep learning in turn is a subset of machine learning, using deep artificial neural networks [12]. Artificial neural networks are models inspired by information processing in biological neural networks with origins in the 1940s [13]. Deep learning models have an input layer (image data), hidden layer(s) (a deep model have several hidden layers) and an output layer (predictions). As information passes through the hidden layers of the model, which include nonlinear activation functions, hierarchical representations of complex patterns in the input data can be learned by the model. Optimization of the model parameters is achieved through iterative updating of the parameters with the objective of minimizing a loss function, which compare the actual labels or response values with predictions from the model. The optimization of deep learning models is typically computationally demanding and require large data sets. Deep convolutional neural networks are a particular type of deep learning models that is used for modelling of image data. Whilst previous generations of ML-based models for image analysis have been depending on human engineering of features, that is patterns and structures in the images analysed, deep learning facilitates end-to-end learning, where feature extraction is an intrinsic part of how the model is optimized to learn representations directly from data. Once the model has been optimized on large amount of training data, the learned patterns captured by the model can be applied to provide predictions of responses or labels in previously unseen observations [14]. Depending on development strategy for a new image analysis models, time-consuming manual annotation of the digitalized slide images may be needed by a pathologist to provide labels to train the model, for example to define cancer or metastasis and to avoid artefacts. However, given large training data sets, slide-level labels (e.g. presence of cancer or not) might be sufficient to train high-performing deep learning models [15]. Ultimately, the architecture and properties of deep neural networks facilitate modelling highly complex and nonlinear patterns in data, and in many instances with exceptional performance.

The integration of AI will be a milestone for health care in the next decade, and pathology is right at the focus of this revolution. Examples of potential added value of AI tools is earlier disease detection, more precise and quantitative diagnosis, discovery of new contexts in human biology, and progress on personalized diagnostics and patient care. In some areas of medical imaging, deep learning algorithms have been proved to have diagnostic performance on par with human experts, or even outperforming them. Deep learning analysis of skin lesion images has reached diagnostic accuracy comparable with dermatologists in detecting squamous carcinomas versus benign seborrhoeic keratoses [16]. In another study, deep learning showed ophthalmologist-level achievement on optical coherence tomography images detecting sight-threatening retinal diseases [17]. Moreover, the US Drug and Food Administration (FDA) has approved a deep learning-based autonomous AI diagnostic system to detect diabetic retinopathy on the images of

retinal fundus [18]. In this review, we discuss the most recent developments and challenges in digital image analysis and the impact of artificial intelligence on diagnostic pathology.

Why digital image analysis in pathology?

Historically, diagnostic pathology has been performed by microscopic evaluation of tissue sections or biopsies on glass slides. The digitalization of microscopic images allows quantitative machine-based image analysis (Fig. 1) and could demonstrate to be clinically valuable as a tool to precisely detect disease and predict patient outcome. Innovative digital image analysis methodologies to improve therapy-response prediction and outcome prognostication have the highest value in the clinicopathological setting. Present classification of biomarkers, based on pathologist's visual assessment, is subjective to some degree. Moreover, visual evaluation and manually counting of cells hinders reproducibility. Although there is controversy about how image analysis should be implemented in the clinical setting [19], computer-based image analysis is bound to increase reproducibility. Emergent examples of urgent need for quantitative histopathology comprise proliferation scoring based on Ki67-immunohistochemistry [20], tumour-infiltrating lymphocytes (TILs) [21] in breast and other cancers and invasive cancer detection.

Ki67 is currently one of the most encouraging although yet controversial biomarker in breast cancer [22]. Despite the promise of Ki67 as a prognostic and/or predictive tool, controversy exists regarding its applied methodology in clinical practice. Therefore, there is an urgent demand for standardized methodology and reproducible scoring methods of Ki67 proliferation index. To overcome this struggle, the International Ki67 in Breast Cancer Working Group (IKWG) has introduced a recommendation for the use of Ki67 IHC in clinical routine [23]. According to this, factors that primarily affect the IHC results of Ki67 include pre-analytical, analytical, interpretation and scoring, and data analysis steps [23]. Although the IKWG recommendations present a guideline to increase preanalytical and analytical reproducibility, inter-laboratory protocols still showed high variability associated with different sampling, fixation, antigen retrieval, staining and scoring methods [7, 23].

The progression of malignant tumours promotes interaction with other cells in the tumour surroundings including immune cells [24]. Due to changed protein expression by the tumour, the immune cells can identify the tumour cells and initiate an immune response [25, 26]. Several studies have demonstrated that assessment of TILs has clinical significance in breast cancer, nonsmall cell lung cancer and melanoma [21, 27–29]. Immune-checkpoint inhibitors are becoming a significant part of treatment for several tumour types. However, its clinical success relies on numerous factors like the expression of immunomodulatory ligands (e.g. PD-L1 [30]), tumour mutational burden and TILs [31, 32]. Visual scoring of TILs has been established for many years, but has not seen wide implementation in clinical decision-making due to insufficiency of standardization between institutions and concerns regarding reproducibility between pathologists [33]. The International TIL working group has undertaken efforts on the standardization of TILs scoring and proposed a guideline for pathologists how to manually assess TILs on HE slides [34, 35]. Although the clinical potential of the this visual TIL scoring guideline has been demonstrated in international ring studies [36], reproducibility will likely remain an issue [37].

Breast cancer grade is one of the most robust prognostic markers to categorize patients into groups with good outcome (grade 1) and with poor outcome and high risk of death (grade 3) [38]. However, Nottingham modification of the Scarff–Bloom–Richardson grading is subject to high inter-pathologist variability [39, 40]. For patients with prostate cancer, the Gleason score is one of the strongest prognostic factors, defining treatment independent of the stage [41]. However, Gleason scoring based on microscopic evaluation of prostate cancer morphology is not only tedious, but also subject to high intra/inter-observe variabilities [42].

Digital image analysis solutions in pathology

Recently, FDA approved the first whole slide imaging system for digital pathology that marks a new era of digital image analysis in pathology [43]. There is currently also a rising interest and competition with respect to digital image analysis solutions for clinical applications. Pathology is an image-related discipline, primarily with the bright-field microscope as the major working platform for

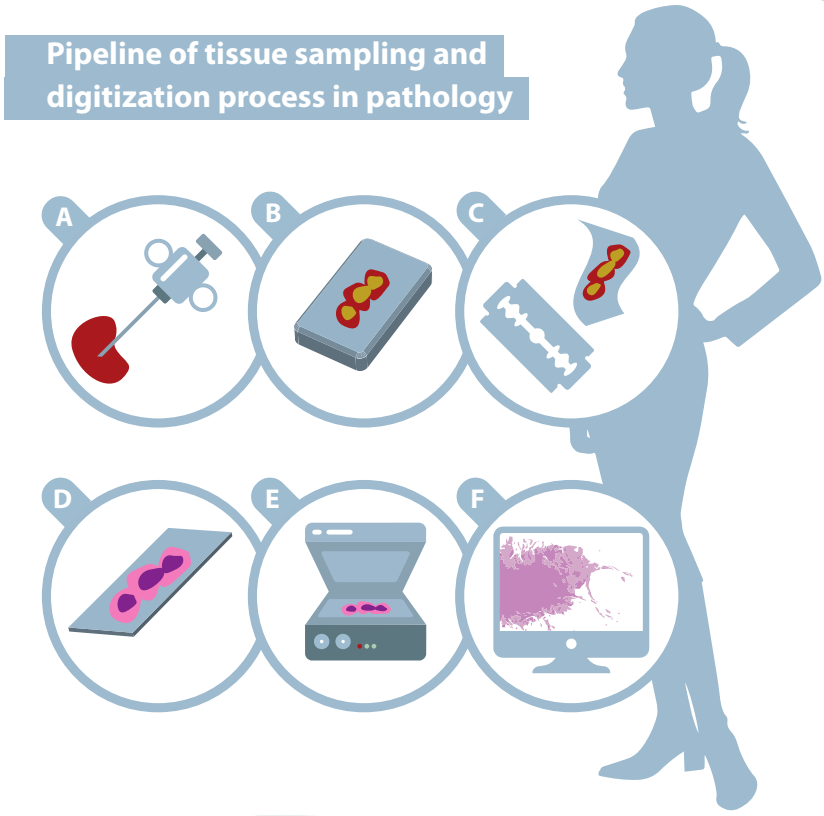

Fig. 1 *Pipeline of tissue sampling and digitization process in pathology. After the biopsy was taken from the patient (a), a tissue block is made preceded by fixation and paraffin embedding (b). After the tissue block was cut (c), the section is put on a glass slide followed by special staining (d). Then, the stained slide is placed in special slide scanner (e), resulting a digitized tissue slide (f). [Correction added on 14 April 2020, after first online publication: Figure 1 has been corrected in this current version.]*

tissue representation. Several digital image analysis platforms have been developed to support the pathologist's assessment of digitized slides. Such applications aim to increase diagnostic accuracy, reliability, reproducibility and efficiency by enabling quantitative image analysis. However, digital image analysis for histopathology has been available for decades. During this time, methods have been developed to decrease variability in image quality, for example colour standardization, spatial filtering, denoising or enhancement [44]. Serious efforts have been made to automate segmentation of cells and cell nuclei through using active contour models [45]. This introduced a universal segmentation method that fits a deformable shape model to the image [46–48]. Mitosis detection has been also in focus in research [49]. For a comprehensive review of object detection and segmentation, see Gurcan et al. and Veta et al. [50, 51]. In this review, we do not attempt to describe all the previous methods for image analysis in histopathology, but to give the current state of DIA in pathology.

Several DIA platforms with various concepts are available for quantitative biomarker evaluation. In the following section, we briefly summarize the most commonly used DIA platforms in pathology. Ventana Companion Algorithm image analysis software is CE and US IVD approved platform for Roche IHC assays in breast pathology to evaluate breast panel biomarkers (ER, PgR, HER2, Ki67 and P53) and to grant an integrated platform including antibody assays. In 2014, AstraZeneca obtained the imaging and data analysis technology firm Definiens, and incorporated the Tissue Phenomics® software that has been introduced to clinical programs in immune-oncology and to support biomarker identification. The above-mentioned applications require a whole slide scanner to function, whereas the Aperio Digital Pathology (Leica Biosystems, Nussloch, Germany) platform integrated a digital microscope with image analysis software. HALO (Indica laboratories) offers IHC and fluorescence modules of quantitative tissue analysis designed mainly for research. QuantCenter is the framework for 3DHISTECH image analysis

applications, designed for the digital whole slide quantification process. QuantCenter provides several modules for tissue classification, IHC quantification and molecular pathology. Cognition Master Professional Suite platform by VMscope can be integrated into laboratory information management system and offers modules for scoring of Ki67, ER, PgR, CD3/4/8/15/20, TIL and vascular stenosis. TissueGnostics analysis software (Vienna, Austria) provides image analysis solutions for clinical and research evaluation of biomarkers in breast cancer. Visiopharm® (Hoersholm, Denmark) Virtual Dual Staining (VDS) method is aligns a pancytokeratin stained consecutive section of the tumour with the IHC stained biomarker to be investigated that allows automated identification of tumour regions [52–54]. The above-mentioned platforms are all commercially available; however, there are also open-source platforms for digital image analysis in pathology. Amongst the first freely accessible tools for image analysis that included morphological parameters is ImageJ, published in 1997, developed by the National Institute of Health (Bethesda, Maryland, USA), is widely used in biomedical image analysis [55]. CellProfiler software was published in 2006 [56] and provides supervised machine learning-based classification to perform imaging-based diagnoses. Another open-source platform is QuPath, which has a special focus on digital pathology and whole slide image analysis [5, 57]. Although published in 2017, there are already more than 122 citations in peer-reviewed publications (as of July 2019), indicating its impact in the field. The software offers unsupervised machine learning-based cell detection and supervised classification of whole slide images, tumour identification and high-throughput biomarker evaluation [5, 57].

Hitherto, most of the studies focusing on digital image analysis has attempted to quantitatively score IHC stained biomarkers. DIA has demonstrated outstanding reproducibility, but studies have mainly been limited to few individual biomarkers or cohorts with low number of cases [53, 58–60]. Furthermore, benchmarking against stronger ground truth variables, including gene expression assays, or outcome data, could be expected to be superior when comparing performance and reproducibility between conventional and digital scoring [52]. Automated image analysis has been used to evaluate numerous biomarkers, including HER2, aiming to decrease additional ISH analysis in HER2 equivocal cases [60, 61] and significantly increasing inter- and intra-observer reproducibility [62, 63]. In respect of TILs, a recently published study demonstrated that automated TIL scoring has a robust and independent prognostic potential in melanoma, whereas pathologist's TIL scoring did not reach statistical significance [64]. DIA systems are now suitable to score Ki67, and numerous studies have been focused on comparing pathologists' visual evaluation with machine's scoring [20, 52, 65–67]. Automated image analysis can be used as screening tool by analysing cytokeratin-stained lymph node sections to eliminate metastasis-free samples with 100% sensitivity [68]. Moreover, major compression and scaling of large digitized whole slides can be achieved without compromising image analysis of biomarker expression [69].

Although image analysis of IHC has went beyond human capability to quantify expression, such DIA systems have not changed pathology daily practice. Potential explanation is that different platforms have unique algorithms to detect and classify objects (cells and tissue compartments) and handle staining intensities. Furthermore, there is very limited number of studies comparing different DIA platforms. The Aperio Digital Pathology operator-supervised system has been compared with the fully automated Definiens Tissue Studio software for scoring ER and PgR expression in breast cancer, showing good correlation between the two platforms [70].

In a recent study, the between-platform concordance was tested in Ki67 scoring between two DIA systems using VDS [71]. Consecutive sections were stained for cytokeratin (CK) 8/18 and Ki67. Then, the authors digitally aligned the corresponding slides to score Ki67 in the CK-positive tumour regions. The authors showed high agreement between the two DIA platform using VDS. Cell detection performance was compared between two platforms in another detailed study. The authors built a DIA algorithm to segment cell nuclei in breast cancer stained with several IHC and FISH markers [72]. The authors compared the sensitivity and positive predictive values (PPV) of the new algorithm and other DIA platforms in cell nuclei segmentation applying pathologist's nuclear marking as a ground truth. Although it was demonstrated that the PPV values ranged between 87 and 94% amongst the different DIA systems, the between-platform reproducibility in Ki67 scoring was not investigated [72].

Although it has long been conceded that detection of Ki67-positive tumour cells might have prognostic and predictive potential in breast cancer [73–76], it has not been widely used in clinical breast cancer management. This is primarily a consequence of insufficient reproducibility in Ki67 scoring across laboratories. Therefore, the IKWG has been investigating DIA in Ki67 scoring and published two comparison studies of different DIA platforms in 2019 [77, 78]. They found that automated DIA assessment of Ki67 has performed similar inter-laboratory reproducibility to that for a rigorously standardized pathologist's visual evaluation of Ki67 [78]. They also demonstrated a very high reproducibility both intra- and inter-DIA platforms, including one open access DIA software (QuPath) [77]. Furthermore, the investigated platforms have very similar prognostic potential in breast cancer-specific overall survival [77].

The rapid emergence of image analysis solutions and integrated platforms for histopathological diagnostics will most likely persist for the upcoming years, resulting close competition amongst companies. However, in order to DIA platforms change patient care, clinical utility must be validated prospectively on routine samples in pathology departments. The emergence of deep learning algorithms (e.g. deep convolutional neural networks) may further facilitate the adoption of digital pathology technologies in daily practice.

Deep learning and its application in pathology

Recent groundbreaking results in AI holds a promise to significantly alter the way we diagnose and stratify cancer in pathology. Deep learning techniques represent a milestone in this transformation, as the application of deep neural network models already are behind several breakthroughs addressing key current issues of histopathology. Several types of deep neural networks exist, whilst convolutional neural network (CNN) [79] is the most commonly used in pathology image analysis [80]. A typical CNN is composed of an input layer, a task-specific output layer and multiple hidden layers. Each hidden layer consists of a number of convolutional filters (parameters) that one by one apply the same local transformations at various locations of its input image [14]. Since the parameters are shared as they are applied locally across the image, an efficient parameterization of the CNN model can be achieved. Typical implementations of CNN models provide a degree of translational invariance, that is allowing that detected objects or patterns can occur at any location in the image. Pooling layers are typically included between convolutional layers to down sample the intermediate outputs (feature maps) from the convolutional functions. The convolutional layers are followed by fully connected layers that flatten the output from convolutional layers and generate the final representations that feed into the output layer [2, 14]. Each neuron in CNN computes an output value by applying a vector of weights and a biases (parameters) to the input values coming from the previous layer. The optimization (training) of a CNN model proceeds by iteratively adjusting these biases and weights in order to minimize a loss function. One advantage of CNNs compared to other image classification algorithms is that it facilitates end-to-end learning. This means that CNN learns the filters (parameters) and representations, which are hand-engineered in conventional algorithms [2]. Another major advantage of CNNs is the flexibility and high capacity of these models to learn patterns in image data, which currently represents state-of-the-art in image analysis and classification and consistently outperforms previous generations of image analysis methodologies [80]. For detailed review of deep learning algorithms, refer to [2, 14, 80]. A plethora of deep learning architectures have been applied in pathology focused research, and several types of modelling aims have been pursued. The recently published studies presenting deep learning applications in pathology aimed to either (i) predict routine diagnostic features used in pathology practice (e g. disease vs normal tissue, define tumour grade and distinguish cancer types) or (ii). Identify novel insights into disease (Fig. 2). Recently published studies attempting to exploit properties of deep neural networks to assess histomorphological features in heamatoxylin–eosin slides (HE). Since deep neural network models require large training and multiple validation sets, recent applications have been focused on the most common types of cancers, namely breast, prostate and lung cancer. However, recent advances in image analysis have also been applied to other malignant tumours such as melanoma [81], pancreatic neuroendocrine tumours [82], ovarian cancer [83], cervical cancer [84] and glioma [85]. Table 1 summarizes the studies reviewed in our paper.

Deep learning in breast cancer pathology

The CAMYLEON16 challenge was the first major challenge on computer-aided diagnosis in

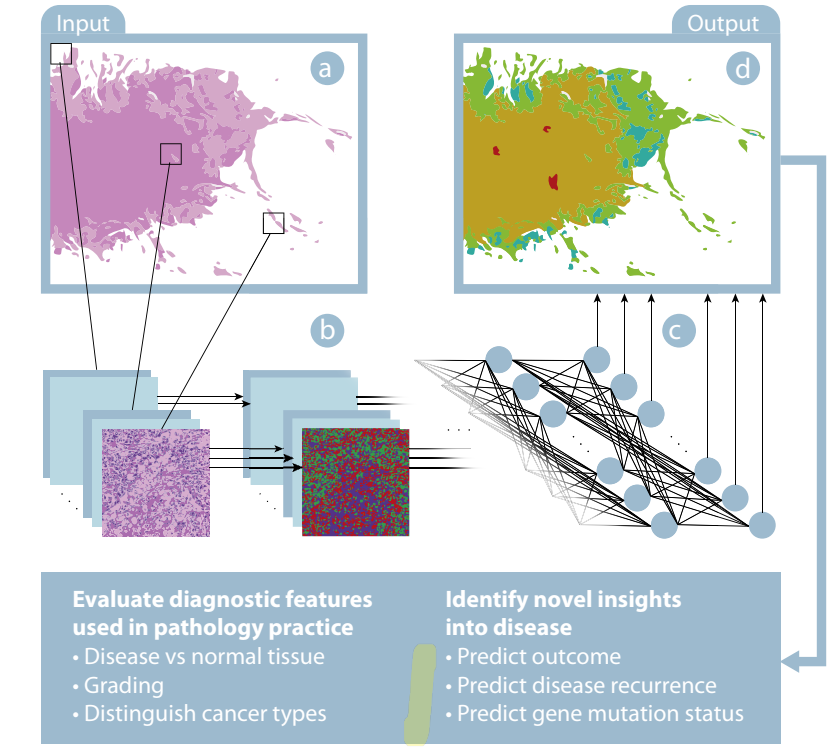

Fig. 2 *Deep learning in Pathology. HE image is divided into patches (a). Convolutional neural network (CNN) typically consists of several convolutional layers; each of these layers applies the same local transformations at various locations of its input image (b). Convolutional layers are followed by fully connected layers that connect every neuron in one layer to every neuron in another layer (c). Each neuron in CNN computes an output value by applying a vector of weights and a bias to the input values coming from the previous layer. Based on this, CNN generates a new representation of the image, with a significant number of training instances, CNN can be used to either evaluate the routine diagnostic features used in pathology practice or identify novel*

histopathology using whole slide images [86]. The data contained HE images from sentinel lymph nodes of breast cancer patients, with the task to identify breast cancer metastasis. This study demonstrated that deep learning algorithms can reach comparable detection of breast cancer micrometastases on HE slides of lymph nodes compared to a pathologist without time-constrained limitation. During the time-constrained exercise to simulate pathology daily routine, deep learning algorithms outperformed the panel of 11 pathologists in detecting micrometastases [86]. Although this study is a milestone in computational pathology, its clinical transition is limited by the current gold standard IHC detection of lymph node metastasis in daily practice that is widely available and can be easily performed with outstanding performance [87]. However, there are other fields in breast pathology, where deep learning can potentially exceed the clinical utility of current ground truth. Romo-Bucheli et al. developed a CNN to identify tubules of breast cancer [88]. They tested their deep learning model on WSIs of 174 patients and found that the CNN-based tubule formation score was associated with the corresponding Oncotype DX and tumour grade risk categories in ER+ breast cancer. In another study, a deep CNN model for classification was trained and tested on digital images of 2387 HE sections of benign and malignant core biopsies from 882 breast cancer patients [89]. The deep learning models were trained to discriminate benign from malignant biopsies based on the stromal compartment. In the test set of 330 patients, the algorithm reached an AUC of 0.962 on slide-level reporting of malignancy. Furthermore, the probability of predicted tumour-associated stroma was correlated with the grading of ductal carcinoma in situ (DCIS) as tumour-associated stroma probabilities were significantly higher in grade 3 DCIS compared to grade 1. In a recent study by Mercan et al., a deep CNN model was trained on images from 240 breast biopsies in order to distinguish normal tissue, atypia, DCIS and invasive cancer [90]. Three pathologists' consensus report was used as the ground truth whilst the deep learning application was compared to 87 participating pathologists. The specificity (0.80) was the same, whilst the accuracy

Table 1. *List of the studies reviewed in our paper*

Publication	Disease	Task	Number of cases involved in the study	Digital image analysis method	Diagnostic performance
Holten-Rossing et al. [60]	Breast Cancer	HER2 IHC scoring	*N* = 462	Segmentation	Compared to FISH analysis, the algorithm reached 100% sensitivity and 95.5% specificity
Acs et al. [64]	Melanoma	Automated TIL scoring on HE slides	*N* = 621	Segmentation and machine learning	Algorithm showed prognostic potential in three independent cohorts
Holten-Rossing et al. [68]	Breast Cancer	Decision support tool - Automated screening of sentinel lymph node biopsies in breast cancer	*N* = 900 (patient *n* = 135)	Segmentation and machine learning	Algorithm showed 100 sensitivity and could have decreased the workload by 58.2%
Ahern et al. [70]	Breast Cancer	Inter-Platform reproducibility: Comparison of Aperio and Definiens DIA systems in ER and PgR scoring against pathologist classification	*N* = 592	Segmentation and machine learning	AUC ranged 0.90–0.97 for ER and 0.87–0.94 for PgR
Koopman et al. [71]	Breast Cancer	Inter-Platform reproducibility: Comparison of Visiopharm and Halo DIA systems in Ki67 scoring against pathologist classification	*N* = 154	Segmentation with virtual dual staining	Inter-platform agreement was 0.96 (spearman correlation)
Acs et al. [77]	Breast Cancer	Inter-Platform and Inter-Operator reproducibility: Comparison of three DIA systems with machine learning in Ki67 scoring run by four operators	Scoring reproducibly *n* = 30; Clinical validity *n* = 149	Segmentation and machine learning	Inter-platform reproducibility was 0.93 (intra-class correlation), Platforms showed similar prognostic potential

Table 1 (*Continued*)

Publication	Disease	Task	Number of cases involved in the study	Digital image analysis method	Diagnostic performance
Rimm et al. [78]	Breast Cancer	Inter-Laboratory reproducibility amongst 14 laboratories: Using digital image analysis in Ki67 scoring	*N* = 30	Automated machine-based scoring: Different methods as 10 DIA platforms were used	Inter-operator reproducibility was 0.89 (intra-class correlation)
Hekler et al. [81]	Melanoma	Comparison of automated classification of melanoma histological subtypes against pathologists	*N* = 695	Deep learning – CNN	CNN outperformed 11 histopathologists in the classification of histopathological melanoma images (sensitivity, accuracy)
Niazi et al. [82]	Pancreatic neuroendocrine tumour (NET)	Automatically distinguish NET and nontumour regions on Ki67 stained biopsies	*N* = 30	Deep learning – CNN	Model showed 97.8% sensitivity and 88.8% specificity against pathologists' classification
Wu et al. [83]	Ovarian Cancer	Automatically classify different ovarian cancer types on HE slides	*N* = 85	Deep learning – CNN	Accuracy of classification was 78.20%
Zhang et al. [84]	Cervical Cancer	Automatically classify cervical cells into abnormal and normal categories on Pap smear and liquid-based cytology	*N* = 1906 (cell count)	Deep learning – CNN	Classification accuracy and specificity was 98.3%, AUC was 0.99
Ertosun [85]	Glioma	Automated grading of gliomas	*N* = 22	Deep learning – CNN	Model achieved a classification of 96% accuracy in differentiating lower grade glioma and glioblastoma multiforme
Bejnordi et al. [86]	Breast Cancer	Automatically detect breast cancer metastasis in lymph nodes	*N* = 399	Several DIA methods applied,	The best algorithm reached better classification (AUC

Table 1 (*Continued*)

Publication	Disease	Task	Number of cases involved in the study	Digital image analysis method	Diagnostic performance
				mostly deep learning -CNN	0.994) than pathologists (AUC 0.810)
Romo-Bucheli [88]	Breast Cancer	Automatically identify tubules of breast cancer	*N* = 174	Deep learning – CNN	The tubule formation score was significantly correlated (AUC 0.76) with the corresponding Oncotype DX categories
Bejnordi et al. [89]	Breast Cancer	Automatically discriminate benign from malignant biopsies based on the stromal compartment	*N* = 2387	Deep learning – CNN	Model reached an AUC of 0.962 on slide-level reporting of malignancy
Mercan et al. [90]	Breast Cancer	Automatically distinguish normal tissue, atypia, DCIS and invasive cancer	*N* = 240	Deep learning – CNN	Model reached sensitivity (89%) and specificity (80%) comparable with 87 pathologists
Campanella et al. [15]	Prostate Cancer, Breast Cancer, Basal Cell Carcinoma (Skin)	Automated cancer detection	*N* = 44 732	Multiple instance learning-based deep learning	Model showed an AUC greater than 0.98 for all cancer types
Campanella et al. [91]	Prostate Cancer	Automated cancer detection	*N* = 12 160	Multiple instance learning-based deep learning	Model showed an AUC of 0.98 in slide-level cancer detection
Litjens et al. [92]	Prostate Cancer	Automated cancer detection	*N* = 225	Deep learning – CNN	Cancer likelihood map achieved an AUC of 0.99 on slide-level detection
Arvaniti et al. [93]	Prostate Cancer	Automated Gleason grading	*N* = 886	Deep learning – CNN	Model showed comparable inter-observer agreement (kappa = 0.71 and 0.75) with that of occurred between the 2 ground truth pathologists (kappa = 0.71).

Table 1 (*Continued*)

Publication	Disease	Task	Number of cases involved in the study	Digital image analysis method	Diagnostic performance
Nagpal et al. [94]	Prostate Cancer	Automated Gleason grading	*N* = 1557	Deep learning – CNN	Model achieved higher diagnostic accuracy (0.70) compared to the mean accuracy amongst 29 pathologists (0.61)
Ström et al. [95]	Prostate Cancer	Automated Gleason grading	*N* = 6682	Deep learning – CNN	Model showed comparable agreement (kappa = 0.62) that occurred amongst 23 human urological pathologists
Coudray et al. [96]	Lung Cancer	Automated classification of normal lung, lung adenocarcinoma (LUAD) and squamous cell carcinoma (LUSC)	*N* = 1974	Deep learning – CNN	Model showed high performance (AUC: 0.97) comparable to that of a pathologist
Nirschl et al. [97]	Heart Failure	Automatically detect clinical heart failure from HE stained endomyocardial biopsies	*N* = 209	Deep learning – CNN	Model showed an AUC of 0.97 in detecting heart failure outperforming the two pathologists' readings (AUC: 0.75)
Wei et al. [98]	Coeliac Disease	Automated classification of coeliac disease, nonspecific duodenitis and normal tissue on HE stained duodenal biopsy	*N* = 1230	Deep learning – CNN	Model achieved slide-level AUC greater than 0.95 for all the three diagnostic classes
Wang et al. [99]	Lung Cancer	Predict recurrence in early-stage nonsmall cell lung cancer (NSCLC) from HE stained TMA slides	*N* = 305	Deep learning with segmentation supported by quadratic discriminant	Model showed accuracy of 82% and 75% for prediction of recurrence and it was also proved to be an independent prognostic factor

Table 1 (*Continued*)

Publication	Disease	Task	Number of cases involved in the study	Digital image analysis method	Diagnostic performance
				analysis (QDA), linear discriminant analysis (LDA), and support vector machine (SVM)	
Kulkarni et al. [100]	Melanoma	Predict disease-specific survival from HE slides	*N* = 263	Deep learning – CNN with segmentation	Model achieved an AUC of 0.880 and 0.905 in two independent cohorts
Mobadersany et al. [101]	Glioma	Predict survival	*N* = 1061	Deep learning – CNN integrated with a Cox proportional hazards model	Model achieved higher prognostic accuracy (median c index 0.801) than the current WHO paradigm based on genomic classification and histologic grading (median c index 0.774)
Saltz et al. [102]	13 cancer types	Mapping TIL patterns and correlate it with molecular subtypes and outcome	*N* = 5202	Deep learning – CNN	Model predicted TIL patterns are differently related to survival amongst different tumour types
Schaumberg et al. [103]	Prostate Cancer	Predict molecular profile: Detect SPOP mutation status on HE slides	*N* = 329	Deep learning – CNN	Model achieved an AUC of 0.74 and 0.86 in two independent cohorts
Coudray et al. [96]	Lung Cancer	Predict molecular profile: Detect KRAS, FAT1, TP53, SETBP1, EGFR, and STK11 mutation status on HE slides	*N* = 567	Deep learning – CNN	Slide-level AUC ranged between 0.733 and 0.856

Table 1 (*Continued*)

Publication	Disease	Task	Number of cases involved in the study	Digital image analysis method	Diagnostic performance
Kather et al. [104]	Gastric and Colorectal Cancer	Predict molecular profile: Detect microsatellite instability mutation status on HE stained FFPE and frozen sections	$N = 1616$	Deep learning – CNN	Patient level AUCs ranged between 0.69 and 0.84 in five independent cohorts

(0.85) and the sensitivity (0.89) were superior to that of pathologists (0.82, 0.80, 0.70, respectively) for the invasiveness classification.

Campanella et al. recently presented a decision support system for pathology [15]. They implemented a multiple instance learning-based deep learning framework to detect cancer. The model(s) were trained and validated on an extensive data set as follows: breast metastasis to lymph nodes data set of 9894 slides, a skin data set of 9962 slides and a prostate core biopsy data set with 24 859 slides. The performance of the application to detect breast cancer (in lymph node), basal cell carcinoma and prostate cancer achieved an AUC> 0.98 for all cancer types. The clinical perspective is that this application would allow pathologists to exclude 65–75% of slides (with 100% sensitivity) in daily practice. This study also demonstrated that patient level reported diagnoses can be used as labels for training slides. Thus, they were able to avoid time-consuming pixel-wise manual annotations.

Deep learning in prostate cancer pathology

Serious efforts have been made to adopt deep learning in prostate cancer pathology. In one study, 12 160 prostate needle biopsy images were collected to evaluate multiple instance learning algorithm aiming to detect invasive prostate cancer [91]. In the test set of 1824 slides, the slide-level cancer detection of the deep learning algorithm achieved an AUC of 0.98. In another study, the CNN model was applied on 225 slides to detect prostate cancer [92]. The developed CNN model produces a cancer likelihood map based on cancer likelihood per pixel that achieved an AUC of 0.99 on slide-level detection of prostate cancer. The automated Gleason grading is one of the most active fields in computational pathology. Arvaniti et al. [93] used 641 TMA images of prostate cancer to train CNN algorithm for automated Gleason scoring. In the test TMA cohort of 245 patients, the CNN application showed comparable inter-observer agreement with that of occurred between the 2 ground truth pathologists. Furthermore, the CNN model's Gleason score assignments significantly stratified patients into groups with distinct disease-specific survival. Moreover, this prognostic potential was superior to that of pathologists scoring. In another study, deep learning model was trained on 1226 HE whole slides from prostatectomies [94]. On the validation set of 331 slides, the deep learning application achieved higher

diagnostic accuracy (0.70) compared to the mean accuracy amongst 29 pathologists (0.61) against the reference standard Gleason scores. In the study of Ström et al., [95] 6682 prostate needle biopsy images were collected to develop CNN models to detect presence of prostate cancer and assignment of Gleason grade. In the independent test set of 1631 slides, the slide-level cancer detection of the deep learning model achieved an AUC of 0.997 at the biopsy level. Concordance between the model-based assignment of Gleason grade was evaluated on 87 biopsies that were also graded by 23 human urological pathologists and achieved a pair-wise Cohen's Kappa 0.62, which was in the range of what was observed between human assessors [95].

Deep learning in lung cancer pathology

Attempts have been made to adopt the advantages of deep learning in lung cancer pathology.

A study conducted by Coudray et al. [96] investigated the potential of CNN that classifies histopathology images into normal lung, lung adenocarcinoma (LUAD) and squamous cell carcinoma (LUSC). The authors obtained 1634 whole slide images (1176 tumour tissue and 459 normal) from The Cancer Genome Atlas (TCGA) that they separated into training (70%) validation (15%) and test (15%) set. Their deep learning model showed high performance (AUC: 0.97) comparable to that of a pathologist in separating LUAD, LUSC and normal lung tissue. As a ground truth, TCGA diagnosis was used. Furthermore, their model showed also good performance (AUC: 0.86–0.97) when it was tested on 340 cases as three independent cohorts (98 frozen sections, 140 FFPE sections and 102 biopsy samples). In this case, the diagnosis performed by pathologists were used as gold standard.

Deep learning in noncancer fields of pathology

In the study of Nirschl et al., [97] the authors developed a CNN application to detect clinical heart failure from HE stained endomyocardial biopsies. Biopsy sections from 104 patients were used for training and samples from 105 patients for independent testing. Their CNN model achieved an AUC of 0.97 in detecting heart failure on HE WSI outperforming the two participating pathologists' readings (AUC: 0.75). In another study by Wei et al., [98] a CNN model was trained and tuned on 1018 duodenal biopsy HE images to distinguish coeliac disease, nonspecific duodenitis and normal tissue. Then, the model was tested on an independent cohort of 212 patients' biopsy samples. As a ground truth, three gastrointestinal pathologists' readings were used. The slide-level classification performance of their deep learning application achieved AUC > 0.95 for all the three diagnostic classes.

Besides assessing the routine diagnostic features for pathology practice, deep learning models have been also proposed to identify novel insights into the pathology of diseases.

Deep learning to predict survival outcome based on HE images

Wang and colleagues trained a machine learning model using nuclear orientation, nuclear shape, texture and tumour architecture to predict recurrence in early-stage nonsmall cell lung cancer (NSCLC) from HE stained TMA slides [99]. Their model was validated on two independent early-stage NSCLC cohorts resulting in 82% and 75% accuracy for prediction of recurrence. Moreover, the model's prediction was also proved to be an independent prognostic factor. Although the model was tested on only 235 patients, the concept is compelling. Another very recently published study by Kulkarni et al. [100] demonstrated that a deep learning model can predict prognosis based on standard HE images in early-stage melanoma. In order to detect region of interest for the training of the model, the author applied nuclear segmentation and cell classification using an open-source platform (QuPath). The CNN algorithm was trained on HE images from 108 patients then validated in two independent cohorts encompassing 155 melanoma patients. Their algorithm achieved an AUC of 0.880 and 0.905 in disease-specific survival prediction. Furthermore, it was also demonstrated that the lymphocyte content is the most important factor to predict outcome in melanoma, but immune infiltration on its own did not reach the same prediction accuracy [100]. The study of Mobadersany et al. [101] aimed to predict survival with a CNN-based model in gliomas. The authors used 1061 WSIs with patient follow-up data from TCGA. Their deep learning model is based on a CNN that was integrated with a Cox proportional hazards model to predict patient outcome. It was demonstrated that the predictive performance of the deep learning model is comparable with neuropathologists' histologic grading. Moreover, the authors further extended the application by

integrating corresponding histology images and genomic data into a single unified prediction framework called genomic survival convolutional neural network (GSCNN model). The GSCNN model achieved higher prognostic accuracy than the current WHO paradigm based on genomic classification and histologic grading. As only a small region of each slide was used for training and prediction, and the selection of these regions of interest within each slide required pathologist guidance, further studies are needed to validate clinical utility.

Saltz et al. [102] developed a deep learning model that provides TIL maps derived through computational staining using CNN on HE images. The authors mapped TIL patterns on 5202 slides across 13 cancer types obtained from TCGA and correlated it with molecular subtypes and outcome. Integrated analysis of TIL maps and molecular data demonstrated that the local patterns and overall structural patterns of TILs are differentially represented amongst tumour types, immune subtypes and tumour molecular subtypes. Moreover, it was also demonstrated these patterns are differently related to survival amongst different tumour types.

Deep learning to predict molecular profile based on HE images

More recently, attempts have been made to predict genetic alterations on HE images using deep learning. In the study by Schaumberg et al., [103] the authors built a deep learning model to predict SPOP mutation status on HE images of prostate cancer. The author trained the model on HE images from 177 prostate cancer cases (including 20 SPOP mutant patients) and applied it on a validation cohort of 152 patients (with 19 SPOP mutant patients). As SPOP is a relatively rare genetic variant, the authors addressed the data imbalance by using a class-balanced stratified-sampling ensemble approach. Their model achieved an AUC of 0.74 and AUC of 0.86 in the two cohorts, respectively. In the study of Coudray et al. [96] that was discussed above, the authors also aimed to predict the most commonly mutated genes in lung adenocarcinoma based on both frozen sections and HE stained FFPE slides. They demonstrated that their CNN model was capable to predict the mutations of six genes (KRAS, FAT1, TP53, SETBP1, EGFR and STK11) with a slide-level AUC range between 0.733 and 0.856 on the input HE images. In another very recent study by Kather et al., [104] the developed deep learning model could predict microsatellite instability directly from HE histology images of gastric (STAD) and colorectal cancer (CRC). The patient level AUCs ranged between 0.69 and 0.84 in five independent cohorts of STAD and CRC encompassing totally 1616 patients using both FFPE and frozen sections (training set n = 1053; validation set n = 563).

Limitations and future perspectives

The present studies illustrate that artificial intelligence has opened doors to technological advances in pathology. Although current results have shown convincingly that in some tasks AI can match the performance of human experts, AI still entails limitations and there are numerous challenges remaining. One of the major concerns posing barrier to clinical adoption of deep learning algorithms is challenges associated with interpretation and understanding of how the complex AI model arrives at its decisions, sometimes referred to as the 'black box' problem. Explainable AI [105] and interpretable machine learning methods are currently a highly active field of research, solutions that offers various degrees of interpretability of deep learning models are already emerging, and we anticipate that that the problem of interpretability will be mitigated, at least in part. Model interpretation might also reveal new hallmarks of a disease, such as the histologic presence of oedema in gliomas that has not been previously recognized as an unfavourable marker, but was detected by AI [101]. It is crucial to establish explainable and interpretable machine learning methods for clinical practice, this would address some of the criticism raised by the medical community. On the other hand, medical practitioners also need to accept to some of these limitations once AI meets all requirements of clinical utility. Another important question is the generalization of AI models and medical decision support tools. Recent results have demonstrated that current AI models, when trained on too small data sets, even using meticulous, pixel-wise labels can present a 20% drop of performance when tested on independent data sets [15]. In order for models to have good generalization properties, the training data have to include a broad and representative sample of biological and morphological variability of the disease, as well as the technical variability introduced in the preanalytical and analytical processes in histopathology, and in the image acquisition process. Challenges relating to technical variability can be addressed either by

standardizing and tightly controlling the process, preprocessing image data to minimize effects of technical variability, or by trying to make the models robust to technical variability. Training or fine-tuning the deep learning model on large and diverse data sets might to a degree reduce the generalization error.

Any new test to be implemented into clinical practice is subject to regulation. The new conformité Europeéne – in vitro diagnostic device regulation (CE-IVDR) from 2022 will significantly affect the European laboratories, which is going to require further clinical evidence defined by Notified Bodies in addition to the existing requirements of self-validation and certification route. It is important to new AI tools shall undergo CE-IVD certification to avoid risk related to potentially nonreproducible laboratory-specific machine learning methods. In order to aid clinical translation, a roadmap and regulatory framework towards routine use of artificial intelligence in pathology have been published [10, 106]. However, we expect the clinical uptake might be slowly evolving as (i) costs for setting up digital slide scanner, image storage, maintenance contracts, image analysis software and IT support systems are substantial; (ii) AI applications have to be demonstrated to be robust and safe in a large population representative and blinded cohorts with detailed clinical follow-up, and also validated prospectively on consecutive cases in a pathology department over a set period of time; (iii) Furthermore, defining the minimal level of performance that AI models would have to achieve for pathologists to accept using them is an issue that has not been addressed yet.

The concept that deep learning-based image analysis can predict mutational status in cancer based on the image of HE stained section is very promising. Whilst there is currently much optimism that AI can predict even molecular subtypes of cancer, the molecular targets that are predicted the best still have relatively low sensitivity and specificity compared to up-to-date molecular testing applied in clinical practice. However, if we consider the increasing performance and outstanding cost-effectiveness potential of deep learning in pathological image classification, this raises several questions: Can AI applications replace some of the expensive molecular tests to screen cancer patients and stratify cancer molecular phenotypes or even predict the probability of response to therapy? To what extent are histomorphological features extracted by AI algorithms associated with proteomics, genomics and molecular signalling pathways? It would also be worth to investigate the performance of deep learning algorithms if the training would integrate radiological and pathological images, as well as molecular profiling data. We are convinced that many of these questions will be addressed in numerous studies over the next few years.

Concluding remarks

As the need for personalized cancer care increases, we face an urgent demand for more accurate biomarker evaluation and more quantitative histopathologic cancer diagnosis to aid and improve therapy decisions. Pathologists need to be equipped with new methodology and tools to deliver the needed diagnostic sensitivity and specificity, and it now seems certain that artificial intelligence is the next step towards precision pathology.

Acknowledgements

Our research is supported by grants from the Swedish Cancer fund, Stockholm County Council, the Swedish Research Council, Swedish Breast Cancer Association and the Stockholm Cancer Society.

Conflict of interest

J.H; advisory boards at AstraZeneca, Roche, Novartis, MSD. Speaker honoraria and travel support from Roche.

Statement of author contributions

BA, MR and JH contributed equally to the conception and design, drafting and critical revision of the manuscript.

Ethical approval

Not applicable.

References

1 Tan D, Lynch HT. *Principles of Molecular Diagnostics and Personalized Cancer Medicine*. Philadelphia: Lippincott Williams & Wilkins, 2012.

2 Robertson S, Azizpour H, Smith K *et al.* Digital image analysis in breast pathology-from image processing techniques to artificial intelligence. *Transl Res* 2018; **194:** 19–35.
3 Maxwell P, Salto-Tellez M. Validation of immunocytochemistry as a morphomolecular technique. *Cancer Cytopathol* 2016; **124:** 540–5.
4 Elliott K, McQuaid S, Salto-Tellez M *et al.* Immunohistochemistry should undergo robust validation equivalent to that of molecular diagnostics. *J Clin Pathol* 2015; **68:** 766–70.
5 Bankhead P, Fernandez JA, McArt DG *et al.* Integrated tumor identification and automated scoring minimizes pathologist involvement and provides new insights to key biomarkers in breast cancer. *Lab Invest* 2018; **98:** 15–26.
6 Varga Z, Diebold J, Dommann-Scherrer C *et al.* How reliable is Ki-67 immunohistochemistry in grade 2 breast carcinomas? A QA study of the Swiss Working Group of Breast- and Gynecopathologists. *PLoS ONE* 2012; **7:** e37379.
7 Polley MY, Leung SC, McShane LM *et al.* An international Ki67 reproducibility study. *J Natl Cancer Inst* 2013; **105:** 1897–906.
8 Bengtsson E. The measuring of cell features. *Anal Quant Cytol Histol* 1987; **9:** 212–7.
9 Ong S, Jin X, Sinniah R. Image analysis of tissue sections. *Comput Biol Med* 1996; **26:** 269–79.
10 Colling R, Pitman H, Oien K *et al.* Artificial intelligence in digital pathology: a roadmap to routine use in clinical practice. *J Pathol* 2019; **249:** 143–50.
11 Hamet P, Tremblay J. Artificial intelligence in medicine. *Metabolism* 2017; **69s:** S36–40.
12 LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature* 2015; **521:** 436–44.
13 McCulloch WS, Pitts W. A logical calculus of the ideas immanent in nervous activity. *Bull Mathe Biophys* 1943; **5:** 115–33.
14 Janowczyk A, Madabhushi A. Deep learning for digital pathology image analysis: a comprehensive tutorial with selected use cases. *J Pathol Inform* 2016; **7:** 29.
15 Campanella G, Hanna MG, Geneslaw L *et al.* Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nat Med* 2019; **25:** 1301–9.
16 Esteva A, Kuprel B, Novoa RA *et al.* Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 2017; **542:** 115–8.
17 De Fauw J, Ledsam JR, Romera-Paredes B *et al.* Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nat Med* 2018; **24:** 1342–50.
18 Abramoff MD, Lavin PT, Birch M *et al.* Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. *NPJ Digit Med* 2018; **1:** 39.
19 Coates AS, Winer EP, Goldhirsch A *et al.* Tailoring therapies-improving the management of early breast cancer: St Gallen International Expert Consensus on the Primary Therapy of Early Breast Cancer 2015. *Ann Oncol* 2015; **26:** 1533–46.
20 Klauschen F, Wienert S, Schmitt WD *et al.* Standardized Ki67 diagnostics using automated scoring-clinical validation in the GeparTrio Breast Cancer Study. *Clin Cancer Res* 2015; **21:** 3651–7.
21 Fridman WH, Pages F, Sautes-Fridman C *et al.* The immune contexture in human tumours: impact on clinical outcome. *Nat Rev Cancer* 2012; **12:** 298–306.
22 Kos Z, Dabbs DJ. Biomarker assessment and molecular testing for prognostication in breast cancer. *Histopathology* 2016; **68:** 70–85.
23 Dowsett M, Nielsen TO, A'Hern R *et al.* Assessment of Ki67 in breast cancer: recommendations from the International Ki67 in Breast Cancer working group. *J Natl Cancer Inst* 2011; **103:** 1656–64.
24 Disis ML. Immune regulation of cancer. *J Clin Oncol* 2010; **28:** 4531–8.
25 Coulie PG, Van den Eynde BJ, van der Bruggen P *et al.* Tumour antigens recognized by T lymphocytes: at the core of cancer immunotherapy. *Nat Rev Cancer* 2014; **14:** 135–46.
26 Gajewski TF, Schreiber H, Fu YX. Innate and adaptive immune cells in the tumor microenvironment. *Nat Immunol* 2013; **14:** 1014–22.
27 Mahmoud SM, Paish EC, Powe DG *et al.* Tumor-infiltrating CD8+ lymphocytes predict clinical outcome in breast cancer. *J Clin Oncol* 2011; **29:** 1949–55.
28 Uryvaev A, Passhak M, Hershkovits D *et al.* The role of tumor-infiltrating lymphocytes (TILs) as a predictive biomarker of response to anti-PD1 therapy in patients with metastatic non-small cell lung cancer or metastatic melanoma. *Med Oncol* 2018; **35:** 25.
29 Rakaee M, Kilvaer TK, Dalen SM *et al.* Evaluation of tumor-infiltrating lymphocytes using routine H&E slides predicts patient survival in resected non–small cell lung cancer. *Human Pathol* 2018; **79:** 188–98.
30 Kluger HM, Zito CR, Turcu G *et al.* PD-L1 studies across tumor types, its differential expression and predictive value in patients treated with immune checkpoint inhibitors. *Clin Cancer Res* 2017; **23:** 4270–9.
31 Ingold Heppner B, Untch M, Denkert C *et al.* Tumor-Infiltrating lymphocytes: a predictive and prognostic biomarker in neoadjuvant-treated HER2-positive breast cancer. *Clin Cancer Res* 2016; **22:** 5747–54.
32 Gibney GT, Weiner LM, Atkins MB. Predictive biomarkers for checkpoint inhibitor-based immunotherapy. *Lancet Oncol* 2016; **17:** e542–51.
33 Dieci MV, Radosevic-Robin N, Fineberg S *et al.* Update on tumor-infiltrating lymphocytes (TILs) in breast cancer, including recommendations to assess TILs in residual disease after neoadjuvant therapy and in carcinoma in situ: A report of the International Immuno-Oncology Biomarker Working Group on Breast Cancer. *Semin Cancer Biol* 2018; **52:** 16–25.
34 Hendry S, Salgado R, Gevaert T *et al.* Assessing tumor-infiltrating lymphocytes in solid tumors: a practical review for pathologists and proposal for a standardized method from the International Immunooncology Biomarkers Working Group: Part 1: Assessing the Host Immune Response, TILs in invasive breast carcinoma and ductal carcinoma in situ, metastatic tumor deposits and areas for further research. *Adv Anat Pathol* 2017; **24:** 235–51.
35 Hendry S, Salgado R, Gevaert T *et al.* Assessing tumor-infiltrating lymphocytes in solid tumors: a practical review for pathologists and proposal for a standardized method from the International Immuno-Oncology Biomarkers Working Group: Part 2: TILs in Melanoma, gastrointestinal tract carcinomas, non-small cell lung carcinoma and mesothelioma, endometrial and ovarian carcinomas, squamous cell carcinoma of the head and neck, genitourinary carcinomas, and primary brain tumors. *Adv Anat Pathol* 2017; **24:** 311–35.

36 Denkert C, Wienert S, Poterie A *et al.* Standardized evaluation of tumor-infiltrating lymphocytes in breast cancer: results of the ring studies of the international immuno-oncology biomarker working group. *Mod Pathol* 2016; **29:** 1155–64.

37 Klauschen F, Muller KR, Binder A *et al.* Scoring of tumor-infiltrating lymphocytes: From visual estimation to machine learning. *Semin Cancer Biol* 2018; **52:** 151–7.

38 Schwartz AM, Henson DE, Chen D *et al.* Histologic grade remains a prognostic factor for breast cancer regardless of the number of positive lymph nodes and tumor size: a study of 161 708 cases of breast cancer from the SEER Program. *Arch Pathol Lab Med* 2014; **138:** 1048–52.

39 Dalton LW, Page DL, Dupont WD. Histologic grading of breast carcinoma. A reproducibility study. *Cancer* 1994; **73:** 2765–70.

40 Dalton LW, Pinder SE, Elston CE *et al.* Histologic grading of breast cancer: linkage of patient outcome with level of pathologist agreement. *Mod Pathol* 2000; **13:** 730–5.

41 Epstein JI, Zelefsky MJ, Sjoberg DD *et al.* A contemporary prostate cancer grading system: a validated alternative to the Gleason score. *Eur Urol* 2016; **69:** 428–35.

42 Egevad L, Ahmad AS, Algaba F *et al.* Standardization of Gleason grading among 337 European pathologists. *Histopathology* 2013; **62:** 247–56.

43 Mukhopadhyay S, Feldman MD, Abels E *et al.* Whole slide imaging versus microscopy for primary diagnosis in surgical pathology: a multicenter blinded randomized noninferiority study of 1992 cases (Pivotal Study). *Am J Surg Pathol* 2017; **42:** 39–52.

44 Khan AM, Rajpoot N, Treanor D *et al.* A nonlinear mapping approach to stain normalization in digital histopathology images using image-specific color deconvolution. *IEEE Trans Biomed Eng* 2014; **61:** 1729–38.

45 Kass M, Witkin A, Terzopoulos D. Snakes: active contour models. *Int J Comput Vis* 1988; **1:** 321–31.

46 Ali S, Madabhushi A. Active contour for overlap resolution using watershed based initialization (ACOReW): Applications to histopathology. In: *2011 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*. Piscataway, NJ: IEEE, 2011; 614–7.

47 Ali S, Madabhushi A. An integrated region-, boundary-, shape-based active contour for multiple object overlap resolution in histological imagery. *IEEE Trans Med Imaging* 2012; **31:** 1448–60.

48 Jing J, Wan T, Cao J *et al.* An improved hybrid active contour model for nuclear segmentation on breast cancer histopathology. In: *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2016; 1155–8.

49 Loi S, Haibe-Kains B, Desmedt C *et al.* Definition of clinically distinct molecular subtypes in estrogen receptor-positive breast carcinomas through genomic grade. *J Clin Oncol* 2007; **25:** 1239.

50 Gurcan MN, Boucheron L, Can A *et al.* Histopathological image analysis: A review. *IEEE Rev Biomed Eng* 2009; **2:** 147.

51 Veta M, Pluim JP, Van Diest PJ *et al.* Breast cancer histopathology image analysis: A review. *IEEE Trans Biomed Eng* 2014; **61:** 1400–11.

52 Stalhammar G, Fuentes Martinez N, Lippert M *et al.* Digital image analysis outperforms manual biomarker assessment in breast cancer. *Mod Pathol* 2016; **29:** 318–29.

53 Roge R, Riber-Hansen R, Nielsen S *et al.* Proliferation assessment in breast carcinomas using digital image analysis based on virtual Ki67/cytokeratin double staining. *Breast Cancer Res Treat* 2016; **158:** 11–9.

54 Lykkegaard Andersen N, Brugmann A, Lelkaitis G *et al.* Virtual double staining: a digital approach to immunohistochemical quantification of estrogen receptor protein in breast carcinoma specimens. *Appl Immunohistochem Mol Morphol* 2018; **26:** 620–6.

55 Schneider CA, Rasband WS, Eliceiri KW. NIH Image to ImageJ: 25 years of image analysis. *Nat Methods* 2012; **9:** 671–5.

56 Carpenter AE, Jones TR, Lamprecht MR *et al.* Cell Profiler: image analysis software for identifying and quantifying cell phenotypes. *Genome Biol* 2006; **7:** R100.

57 Bankhead P, Loughrey MB, Fernandez JA *et al.* QuPath: Open source software for digital pathology image analysis. *Sci Rep* 2017; **7:** 16878.

58 Rizzardi AE, Johnson AT, Vogel RI *et al.* Quantitative comparison of immunohistochemical staining measured by digital image analysis versus pathologist visual scoring. *Diagn Pathol* 2012; **7:** 42.

59 Nielsen PS, Riber-Hansen R, Jensen TO *et al.* Proliferation indices of phosphohistone H3 and Ki67: strong prognostic markers in a consecutive cohort with stage I/II melanoma. *Mod Pathol* 2013; **26:** 404–13.

60 Holten-Rossing H, Moller Talman ML, Kristensson M *et al.* Optimizing HER2 assessment in breast cancer: application of automated image analysis. *Breast Cancer Res Treat* 2015; **152:** 367–75.

61 Brugmann A, Eld M, Lelkaitis G *et al.* Digital image analysis of membrane connectivity is a robust measure of HER2 immunostains. *Breast Cancer Res Treat* 2012; **132:** 41–9.

62 Bloom K, Harrington D. Enhanced accuracy and reliability of HER-2/neu immunohistochemical scoring using digital microscopy. *Am J Clin Pathol* 2004; **121:** 620–30.

63 Gavrielides MA, Gallas BD, Lenz P *et al.* Observer variability in the interpretation of HER2/neu immunohistochemical expression with unaided and computer-aided digital microscopy. *Arch Pathol Lab Med* 2011; **135:** 233–42.

64 Acs B, Ahmed FS, Gupta S *et al.* An open source automated tumor infiltrating lymphocyte algorithm for prognosis in melanoma. *Nat Commun* 2019; **10:** 5440.

65 Zhong F, Bi R, Yu B *et al.* A comparison of visual assessment and automated digital image analysis of Ki67 labeling index in breast cancer. *PLoS ONE* 2016; **11:** e0150505.

66 Stalhammar G, Robertson S, Wedlund L *et al.* Digital image analysis of Ki67 in hot spots is superior to both manual Ki67 and mitotic counts in breast cancer. *Histopathology* 2017; **72:** 974–89.

67 Acs B, Madaras L, Kovacs KA *et al.* Reproducibility and prognostic potential of Ki-67 proliferation index when comparing digital-image analysis with standard semi-quantitative evaluation in breast cancer. *Pathol Oncol Res* 2018; **24:** 115–27.

68 Holten-Rossing H, Talman MM, Jylling AMB *et al.* Application of automated image analysis reduces the workload of manual screening of sentinel lymph node biopsies in breast cancer. *Histopathology* 2017; **71:** 866–73.

69 Konsti J, Lundin M, Linder N *et al.* Effect of image compression and scaling on automated scoring of

immunohistochemical stainings and segmentation of tumor epithelium. *Diagn Pathol* 2012; **7:** 29.
70 Ahern TP, Beck AH, Rosner BA *et al.* Continuous measurement of breast tumour hormone receptor expression: a comparison of two computational pathology platforms. *J Clin Pathol* 2017; **70:** 428–34.
71 Koopman T, Buikema HJ, Hollema H *et al.* Digital image analysis of Ki67 proliferation index in breast cancer using virtual dual staining on whole tissue sections: clinical validation and inter-platform agreement. *Breast Cancer Res Treat* 2018; **169:** 33–42.
72 Paulik R, Micsik T, Kiszler G *et al.* An optimized image analysis algorithm for detecting nuclear signals in digital whole slides for histopathology. *Cytometry A* 2017; **91:** 595–608.
73 Acs B, Zambo V, Vizkeleti L *et al.* Ki-67 as a controversial predictive and prognostic marker in breast cancer patients treated with neoadjuvant chemotherapy. *Diagn Pathol* 2017; **12:** 20.
74 Brown JR, DiGiovanna MP, Killelea B *et al.* Quantitative assessment Ki-67 score for prediction of response to neoadjuvant chemotherapy in breast cancer. *Lab Invest* 2014; **94:** 98–106.
75 Criscitiello C, Disalvatore D, De Laurentiis M *et al.* High Ki-67 score is indicative of a greater benefit from adjuvant chemotherapy when added to endocrine therapy in luminal B HER2 negative and node-positive breast cancer. *Breast* 2014; **23:** 69–75.
76 Stuart-Harris R, Caldas C, Pinder SE *et al.* Proliferation markers and survival in early breast cancer: a systematic review and meta-analysis of 85 studies in 32,825 patients. *Breast* 2008; **17:** 323–34.
77 Acs B, Pelekanou V, Bai Y *et al.* Ki67 reproducibility using digital image analysis: an inter-platform and inter-operator study. *Lab Invest* 2019; **99:** 107–17.
78 Rimm DL, Leung SCY, McShane LM *et al.* An international multicenter study to evaluate reproducibility of automated scoring for assessment of Ki67 in breast cancer. *Mod Pathol* 2019; **32:** 59–69.
79 LeCun Y, Kavukcuoglu K, Farabet C. Convolutional networks and applications in vision. In: *Proceedings of 2010 IEEE International Symposium on Circuits and Systems*. IEEE, 2010; 253–6.
80 Chang HY, Jung CK, Woo JI *et al.* Artificial Intelligence in Pathology. *J Pathol Transl Med* 2019; **53:** 1–12.
81 Hekler A, Utikal JS, Enk AH *et al.* Deep learning outperformed 11 pathologists in the classification of histopathological melanoma images. *Eur J Cancer* 2019; **118:** 91–6.
82 Niazi MKK, Tavolara TE, Arole V *et al.* Identifying tumor in pancreatic neuroendocrine neoplasms from Ki67 images using transfer learning. *PLoS ONE* 2018; **13:** e0195621.
83 Wu M, Yan C, Liu H *et al.* Automatic classification of ovarian cancer types from cytological images using deep convolutional neural networks. *Biosci Rep* 2018; **38.**
84 Zhang L, Le L, Nogues I *et al.* DeepPap: deep convolutional networks for cervical cell classification. *IEEE J Biomed Health Inform* 2017; **21:** 1633–43.
85 Ertosun MG, Rubin DL. Automated grading of gliomas using deep learning in digital pathology images: A modular approach with ensemble of convolutional neural networks. In: *AMIA Annual Symposium Proceedings*. American Medical Informatics Association, 2015; 1899.
86 Bejnordi B, Veta M, van Diest P *et al.* Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA* 2017; **318:** 1–12.
87 Acs B, Rimm DL. Not just digital pathology, intelligent digital pathology. *JAMA Oncol* 2018; **4:** 403–4.
88 Romo-Bucheli D, Janowczyk A, Gilmore H *et al.* Automated tubule nuclei quantification and correlation with oncotype DX risk categories in ER+ breast cancer whole slide images. *Sci Rep* 2016; **6:** 32706.
89 Ehteshami Bejnordi B, Mullooly M, Pfeiffer RM *et al.* Using deep convolutional neural networks to identify and classify tumor-associated stroma in diagnostic breast biopsies. *Mod Pathol* 2018; **31:** 1502–12.
90 Mercan E, Mehta S, Bartlett J *et al.* Assessment of machine learning of breast pathology structures for automated differentiation of breast cancer and high-risk proliferative lesions. *JAMA Netw Open* 2019; **2:** e198777.
91 Campanella G, Silva VWK, Fuchs TJ. Terabyte-scale deep multiple instance learning for classification and localization in pathology. *arXiv preprint* 2018; arXiv:18050 6983.
92 Litjens G, Sanchez CI, Timofeeva N *et al.* Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. *Sci Rep* 2016; **6:** 26286.
93 Arvaniti E, Fricker KS, Moret M *et al.* Automated Gleason grading of prostate cancer tissue microarrays via deep learning. *Sci Rep* 2018; **8:** 12054.
94 Nagpal K, Foote D, Liu Y *et al.* Development and validation of a deep learning algorithm for improving Gleason scoring of prostate cancer. *NPJ Digit Med* 2019; **2:** 48.
95 Ström P, Kartasalo K, Olsson H *et al.* Pathologist-level grading of prostate biopsies with artificial intelligence. *arXiv preprint* 2019; arXiv:190701368.
96 Coudray N, Ocampo PS, Sakellaropoulos T *et al.* Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. *Nat Med* 2018; **24:** 1559–67.
97 Nirschl JJ, Janowczyk A, Peyster EG *et al.* A deep-learning classifier identifies patients with clinical heart failure using whole-slide images of H&E tissue. *PLoS ONE* 2018; **13:** e0192726.
98 Wei JW, Wei JW, Jackson CR *et al.* Automated detection of celiac disease on duodenal biopsy slides: a deep learning approach. *J Pathol Inform* 2019; **10:** 7.
99 Wang X, Janowczyk A, Zhou Y *et al.* Prediction of recurrence in early stage non-small cell lung cancer using computer extracted nuclear features from digital H&E images. *Sci Rep* 2017; **7:** 13543.
100 Kulkarni PM, Robinson EJ, Sarin Pradhan J *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* 2019: clincanres.1495.2019.
101 Mobadersany P, Yousefi S, Amgad M *et al.* Predicting cancer outcomes from histology and genomics using convolutional networks. *Proc Natl Acad Sci U S A* 2018; **115:** E2970–9.
102 Saltz J, Gupta R, Hou L *et al.* Spatial organization and molecular correlation of tumor-infiltrating lymphocytes using deep learning on pathology images. *Cell Rep* 2018; **23:** 181–193.e187.

103 Schaumberg AJ, Rubin MA, Fuchs TJ. H&E-stained whole slide image deep learning predicts SPOP mutation state in prostate cancer. *BioRxiv* 2018; 064279.
104 Kather JN, Pearson AT, Halama N *et al.* Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. *Nat Med* 2019; **25:** 1054–6.
105 Samek W, Wiegand T, Müller K-R. Explainable artificial intelligence: understanding, visualizing and interpreting deep learning models. *arXiv Preprint* 2017; arXiv:17080 8296.
106 US Food and Drug Administration. 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. 2019.

Correspondence: Johan Hartman, Department of Oncology and Pathology, Karolinska Institutet, R8:04, CCK, Karolinska University Hospital, 17177 Stockholm, Sweden.
(fax: +4686162895; e-mail: johan.hartman@ki.se).

