Journal of Pathology
J Pathol July 2022; **257:** 430–444
Published online 21 April 2022 in Wiley Online Library
(wileyonlinelibrary.com) DOI: 10.1002/path.5898

INVITED REVIEW

Artificial intelligence to identify genetic alterations in conventional histopathology

Didem Cifci[1], Sebastian Foersch[2] and Jakob Nikolas Kather[1,3,4]*

[1] *Department of Medicine III, University Hospital RWTH Aachen, Aachen, Germany*
[2] *Institute of Pathology, University Medical Center of the Johannes Gutenberg-University, Mainz, Germany*
[3] *Pathology and Data Analytics, Leeds Institute of Medical Research at St James's, University of Leeds, Leeds, UK*
[4] *Medical Oncology, National Center for Tumor Diseases, University Hospital Heidelberg, Heidelberg, Germany*

**Correspondence to: JN Kather, Department of Medicine III, RWTH University Hospital, 52074 Aachen, Germany. E-mail: jakob-nikolas.kather@alumni.dkfz.de*

Abstract

Precision oncology relies on the identification of targetable molecular alterations in tumor tissues. In many tumor types, a limited set of molecular tests is currently part of standard diagnostic workflows. However, universal testing for all targetable alterations, especially rare ones, is limited by the cost and availability of molecular assays. From 2017 to 2021, multiple studies have shown that artificial intelligence (AI) methods can predict the probability of specific genetic alterations directly from conventional hematoxylin and eosin (H&E) tissue slides. Although these methods are currently less accurate than gold standard testing (e.g. immunohistochemistry, polymerase chain reaction or next-generation sequencing), they could be used as pre-screening tools to reduce the workload of genetic analyses. In this systematic literature review, we summarize the state of the art in predicting molecular alterations from H&E using AI. We found that AI methods perform reasonably well across multiple tumor types, although few algorithms have been broadly validated. In addition, we found that genetic alterations in *FGFR*, *IDH*, *PIK3CA*, *BRAF*, *TP53*, and DNA repair pathways are predictable from H&E in multiple tumor types, while many other genetic alterations have rarely been investigated or were only poorly predictable. Finally, we discuss the next steps for the implementation of AI-based surrogate tests in diagnostic workflows.

Keywords: image analysis; biomarker; artificial intelligence; precision oncology

Received 29 December 2021; Revised 9 March 2022; Accepted 23 March 2022

Conflict of interest statement: JNK declared consulting services for Owkin, France and Panakeia, UK, and has received honoraria for scientific talks and participation in advisory boards by MSD, Eisai, and Bayer. DC and SF declared no conflicts of interest.

Background

Histopathology slides as a high-density source of information

Histopathology is the backbone of cancer diagnostics – for almost every patient with a solid tumor, the final diagnosis is made by a pathologist using microscopy. Routine histopathology images of tissue specimens stained with hematoxylin and eosin (H&E) contain an immense amount of useful information. In addition to diagnostic information, this includes standard prognostic information such as tumor differentiation, tumor budding, lymphovascular invasion, vascular invasion, and perineural invasion, among others. Other common prognostic biomarkers evaluated in research studies include the tumor–stroma ratio [1], tumor-infiltrating lymphocytes [2–8], stromal morphology [9], and the presence of necrosis [10]. In addition to these prognostic patterns, specific morphological patterns have been linked to specific genetic alterations in cancer, such as *BRAF* mutations [11] and microsatellite instability (MSI) in colorectal cancer [12], hormone receptor overexpression in breast cancer [13], and *EGFR* mutations in lung cancer [14]. In summary, H&E-stained tissue sections can reflect specific molecular alterations.

Computer-based image analysis in histopathology

As the workload of pathologists increases in both quantity and complexity, they have limited time to devote to individual cases, which tend to be more and more challenging. This holds true for oncologic cases in particular. The impending scarcity of high-quality pathology services is being further aggravated by an aging pathology workforce and a lack of younger medical professionals moving into this field [15]. Computer-based image analysis was proposed decades ago as a useful strategy to address this problem [16]. Early studies used rule-based image analysis [2,17], or so-called 'classical' machine

learning algorithms, including support vector machines [18,19] or random forest classifiers [20,21]. The performance of these image analysis algorithms was massively improved by the advent of deep learning (DL), in particular by deep convolutional neural networks. These networks outperformed conventional image classification methods in non-medical applications in 2012 [22], with medical applications following shortly after [23].

Deep learning and artificial intelligence

DL is part of the vast field of 'artificial intelligence' (AI). While AI has been used to automate board games [24], video games [25], and other complex tasks, in the context of medicine it is mostly used for image analysis. This includes radiology, dermatoscopy, and histopathology image analysis. In prognostication, DL has been shown to outperform some established risk factors in colorectal cancer [26–28], breast cancer [27], lung cancer [29], sarcoma [30], bladder cancer [31], glioma [32], mesothelioma [33], and hepatocellular carcinoma [34,35], among other tumor types. Compared with prognostic patterns in conventional histology slides, morphological patterns reflecting specific molecular alterations are generally weaker. Although typical morphological patterns are part of pathology textbooks for some genetic alterations (e.g. MSI in colorectal cancer [12]), many such links remain unclear. In 2018, a landmark study by Coudray *et al* surprisingly showed that multiple clinically relevant mutations in lung cancer are predictable from digitized H&E slides alone [29]. While the prediction performance in this initial publication was too low for any immediate clinical application, it proved the concept and sparked dozens of follow-up publications [36–38]. Currently, just a little over 3 years after this initial publication, the prediction of molecular alterations from H&E has been shown to yield good results across a number of tumor types in academic studies. In addition to these research studies, molecular profiling from H&E slides is receiving considerable commercial interest. Although industrial implementation typically follows academic publication with a multi-year lag, it is already clear that multiple commercial entities are focusing on this field and are preparing to launch commercial products for molecular subtyping of tumors from digitized H&E slides. For example, researchers affiliated with PathAI (Boston, MA, USA) have performed a study on prediction of homologous recombination deficiency (HRD) from H&E [39]. Researchers from Owkin (New York, NY, USA) have used DL to predict molecular alterations in mesothelioma [40]; researchers from Panakeia (Cambridge, UK) have published a pan-cancer molecular prediction study [41]; and researchers affiliated with Histofy (Birmingham, UK) have used DL to predict molecular alterations in colorectal cancer [42]. These examples are not a complete list of commercial efforts in this field but demonstrate that prediction of molecular alterations directly from H&E slides is seen as a business opportunity by many.

Aim of this work

Prompted by this broad academic and commercial interest in AI-based prediction of genetic alterations from cancer histology, we performed a systematic review of this field. We report and give analytic details of studies published between 2017 and 2021 and which showed that genetic alterations in tumor tissues are predictable using H&E slides with AI-based methods (supplementary material, Figure S1). The search yielded 52 filtered studies, listed in Table 1, which we categorized as follows: mutation, tumor mutation burden (TMB), DNA damage response (MSI/HRD), gene expression, copy number alteration, and prediction of the presence of oncogenic virus (Figure 1; details given in Supplementary materials and methods).

Statistical endpoints

Prediction of genetic alterations with DL is based on the hypothesis that the algorithms will be able to decipher genetically-associated morphological changes in digitized whole-slide images (WSIs) of conventional H&E slides [29,85]. In this review, we report the most commonly used statistical endpoint used in these studies, the area under the receiver operating characteristic curve (AUROC). Receiver operating characteristic (ROC) curves plot the true positive rate (TPR) (sensitivity) against the false positive rate (FPR) (1 − specificity) at every possible threshold value for decision making [90]. AUROC therefore serves as a very useful method to compare the accuracy of different models and is commonly used as a performance evaluation metric in DL applications [90,91]. Of note, there are some drawbacks to using AUROC as the only performance metric. For example, the AUROC is affected by class imbalance [92]. Also, an AUROC does not specify a fixed threshold or operating point, but this is required for clinical application [93]. Therefore, studies should also include additional metrics such as precision–recall curves or F1 scores. Ideally, the metrics used in a study should be pre-defined [94].

Deep learning for prediction of genetic alterations

Prediction of mutations from H&E slides

In their seminal study published in 2018, Coudray *et al* developed a DL-based image analysis method for mutation prediction in non-small lung cancer (NSCLC), where mutations in *STK11*, *EGFR*, *FAT1*, *SETBP1*, *KRAS*, and *TP53* were predicted from histology with patient-level AUROCs of 0.85, 0.75, 0.74, 0.79, 0.81, and 0.67 on the held-out dataset (i.e. the test dataset) [29].

While most efforts to utilize DL to predict mutations have been in tumor types for which there are actionable alterations, such as lung cancer, a broad spectrum of tumors has been analyzed in recent years. In general, it seems that not all tumor types are equally suitable for these types of analyses. Previous studies focused on tumor entities for which (i) lots of training data are

Table 1. The publications that studied the genetic alteration detection by AI and that are used in this review categorized by the tumor/organ type

Tumor type	Citation	Year	Prediction category	Target	External validation AUROC	Technology
Bladder						
	[43]	2019	TMB	TMB	N/A	CNN, MIL
	[44]	2020	Gene expression	DN (double negative), basal, luminal, and luminal p53-like subtypes of bladder cancer	N/A	CNN
	*[45]	2020	TMB	TMB	N/A	SVM classifier with RBF and linear kernels
	[46]	2021	Mutation	*FGFR3*	AUROC = 0.63	CNN
	[47]	2021	Mutation	*FGFR*	N/A	CNN
Brain						
	[48]	2020	Mutation	*IDH1*	N/A	CNN, MIL
	[49]	2021	Mutation	*IDH* (derived from *IDH1* and *IDH2* status)	N/A	CNN
Blood						
	[50]	2021	Mutation	Mutations in cell cycle; cell differentiation; DNA chromatin structure; *RAS* pathway *IDH1*, *IDH2*, *NRAS*, *KRAS*, and spliceosome genes	N/A	VGG16 and Xception CNNs
Breast						
	[51]	2018	Overexpression	ER	N/A	CNN
	[52]	2019	Overexpression	19 biomarkers	N/A	Morphological-based molecular profiling: logistic regression and CNN
	[53]	2020	Overexpression	HER2	AUROC = 0.76	CNN
	[54]	2020	Gene expression	Expression of 250 genes	AUROC = 0.73 (average AUROC of gene prediction as high or low)	DenseNet-121 CNN
	*[55]	2020	DDR	HRD	AUROC = 0.70	CNN, MIL, RNN
	[56]	2021	DDR	HRD	N/A	Momentum contrast, CNN, MIL
	*[57]	2021	DDR	HRD	N/A	SimCLR, CNN, MIL
	*[58]	2021	CNA, mutation	CNA status in *FGFR1*, *EIF4EBP1*, *KAT6A*, *HEY1*, *ZNF217*, and *RAB25*; mutation in *RB1*, *CDH1*, *NF1*, and *NOTCH2*	N/A	CNN
	[59]	2021	Mutation	*gBRCA*	AUROC = 0.77	CNN
Colorectal						
	*[36]	2019	DDR	Mismatch repair deficiency	AUROC = 0.84	CNN
	[60]	2020	DDR	Mismatch repair deficiency	AUROC = 0.85	CNN, MIL
	[61]	2020	DDR	Mismatch repair deficiency	AUROC = 0.96	CNN
	[62]	2021	DDR, mutation	Mismatch repair deficiency; *TP53* and *BRAF* mutations	AUROC = 0.98 (MSI), AUROC = N/A (*TP53* and *BRAF*)	CNN, HoVer-Net
	[63]	2021	Mutation	*APC*, *KRAS*, *PIK3CA*, *SMAD4*, and *TP53*	AUROC = 0.65 (*APC*), AUROC = 0.58 (*KRAS*), AUROC = 0.57 (*PIK3CA*), AUROC = 0.65 (*SMAD4*), AUROC = 0.78 (*TP53*)	CNN
	[64]	2021	DDR	Mismatch repair deficiency	AUROC = 0.97	Inception-v3 CNN
	*[57]	2021	DDR	Mismatch repair deficiency	N/A	SimCLR, CNN, MIL
	[28]	2021	DDR, mutation	Mismatch repair deficiency; *BRAF* and *KRAS* mutations	AUROC = 0.90 (MSI), AUROC = N/A (*BRAF* and *KRAS*)	ShuffleNet CNN
	[65]	2021	Gene expression	Consensus molecular subtype	AUROC = 0.85 (in TCGA dataset), AUROC = 0.85 (on GRAMPIAN dataset)	Inception CNN
	[66]	2021	DDR	Mismatch repair deficiency	AUROC = 0.779	MobileNetV2 CNN

(Continues)

Table 1. Continued

Tumor type	Citation	Year	Prediction category	Target	External validation AUROC	Technology
Endometrial						
	[67]	2021	CNA, DDR	*POLE* ultra-mutated, MSI-high hypermutated, CNV-L, CNV-H subtypes; mutation status of 18 endometrial carcinoma-related genes	N/A	Inception CNN
Gastric						
	[68]	2017	Overexpression	HER2	N/A	CNN
	*[36]	2019	DDR	Mismatch repair deficiency	AUROC = 0.75	CNN
	*[55]	2020	DDR	Mismatch repair deficiency	AUROC = 0.81	CNN, RNN
	[69]	2021	DDR, oncogenic virus	Mismatch repair deficiency; Epstein–Barr virus	AUROC = 0·86 (MSI), AUROC = 0.86 (EBV) (highest AUROC among several external datasets)	ShuffleNet CNN
	[70]	2021	Oncogenic virus	Epstein–Barr virus	N/A	
Head and neck						
	[71]	2019	Oncogenic virus	Epstein–Barr virus and human papilloma virus	AUROC = 0.81 (EBV), AUROC = 0.70 (HPV)	CNN
	[72]	2021	Oncogenic virus	Human papillomavirus	AUROC = 0.8 (on two independent cohorts separately)	U-Net, DenseNet CNN
Kidney						
	[73]	2021	CNA, TMB	CNA status of clinically important genes; TMB	–	CNN
Liver						
	[74]	2019	TMB	TMB	N/A	
	[75]	2020	Mutation	*CTNNB1*, *FMN2*, *TP53*, and *ZFX4*	† AUROC = 0.90 (*CTNNB1*), AUROC = 0.74 (*FMN2*), AUROC = 0.77 (*TP53*), AUROC = 0.72 (*ZFX4*)	Inception CNN
	[76]	2020	Mutation	*ALB*, *CSMD3*, *CTNNB1*, *MUC4*, *OBSCN*, *TP53*, and *RYR2*	† AUROC = 0.73 (*ALB*), AUROC = 0.75 (*CSMD3*), AUROC = 0.63 (*CTNNB1*), AUROC = 0.63 (*MUC4*), AUROC = 0.74 (*OBSCN*), AUROC = 0.69 (*TP53*), AUROC = 0.80 (*RYR2*)	CNN
	*[58]	2021	CNA, mutation	CNA status in *TGFβ2*; mutation in *RB1* and *NF1*	N/A	CNN
Lung						
	[29]	2018	Mutation	*STK11*, *EGFR*, *FAT1*, *SETBP1*, *KRAS*, and *TP53*	N/A	
	[77]	2019	Overexpression	*PD-L1*	N/A	CNN
	[78]	2020	Mutation	*EGFR*	AUROC = 0.72	Deep learning, not specified
	*[45]	2020	TMB	TMB	N/A	SVM classifier with RBF and linear kernels
	[79]	2021	TMB	TMB	N/A	Inception CNN
	*[58]	2021	CNA, mutation	CNA status in *FGFR1*; mutation in *TP53* and *NOTCH2*	N/A	CNN
Ovarian						
	[80]	2021	DDR, mutation	Mismatch repair deficiency; *BRCA1* and *BRCA2* mutations	N/A	CellProfiler
Pan-cancer						
	[37]	2020	CNA, mutation, gene expression	Whole-genome duplications, copy number alterations, driver gene mutations, gene expression	–	PC-CHiP (using CNN)

(Continues)

Table 1. Continued

Tumor type	Citation	Year	Prediction category	Target	External validation AUROC	Technology
	[81]	2020	DDR, gene expression, mutation, overexpression	Genetic variants, oncogenic drivers, molecular subtypes and gene expression signatures, status of hormone receptors	–	CNN
	[82]	2020	Mutation	*TP53*	–	CNN
	[83]	2020	DDR, gene expression	Mismatch repair deficiency; expression of 30 839 (coding/noncoding) genes	–	HE2RNA
	[84]	2021	DDR, gene expression	PD-1, PD-L1, and CTLA-4 (cytotoxic T-lymphocyte-associated protein) expressions; HRD score and TIGIT (T-cell immunoreceptor with Ig and ITIM domains) expression	-	CNN
Prostate						
	[85]	2018	Mutation	*SPOP*	AUROC = 0.86	CNN
Skin						
	[38]	2019	Mutation	*BRAF* and *NRAS*	N/A	Inception CNN
	[86]	2019	Mutation	*BAP1* (BRCA1-associated protein 1)	N/A	DNN
	[87]	2021	Mutation	*BAP1* (BRCA1-associated protein 1)	N/A	CNN, MIL, CLAM
Thyroid						
	[88]	2019	Mutation	$BRAF^{V600E}$ and *RAS*	N/A	Inception CNN
	[89]	2021	Mutation	$BRAF^{V600E}$	AUROC = 0.98	VGG16 CNN, DNN, MIL

The references to these studies are given along with the publication year, the category of genetic alterations in which the study falls, the predicted targets, the performance of the model, and the AI technology being used in each publication.
CNA, copy number alteration; DDR, DNA damage response; TMB, tumor mutation burden.
*Duplicate entries that studied multiple types of cancer. The performances of pan-cancer studies and the studies with more than seven targets are not given due to space limitations.
†Predictions are made with two methods, based on average predicted probability and summarizing the percentage of positively classified patches from the slides. The selected AUROCs reported here are calculated based on the average predicted probability of the predictions.

available, and (ii) clinical management can be informed by molecular alterations (Figure 2). Here, we review the available evidence on the prediction of mutations in various tumor types.

In one of the studies succeeding Coudray *et al*, *FGFR* mutation status was also shown to be predicted in NSCLC, with an AUROC of 0.72 on the held-out dataset by Wang *et al* [78]. In breast cancer, the mutation status of germline *BRCA*, a biomarker that plays an important role in the DNA repair mechanism and genomic stability, was also found to be predictable, with an AUROC of 0.77 on an external dataset [59,95]. Similarly, detection of *FGFR* mutation status, which is a prognostic biomarker in bladder cancer, was reported in two different publications, with AUROCs from 0.63 to 0.76 for within-cohort and external validation [46,47]. Another clinically relevant set of genetic alterations are mutations occurring in isocitrate dehydrogenase (IDH) enzymes, which exist in the majority of lower-grade gliomas and influence therapeutic decisions [96]. Jiang *et al* used a DL-based histopathology image analysis to train a neural network for the classification of mutations in *IDH* (IDH1 and IDH2) in lower-grade (stages 2 and 3) glioma from The Cancer Genome Atlas (TCGA) cohort and validated it on stage 4 gliomas from the same dataset, which achieved high performance (AUROC of 0.81) [49]. When the whole TCGA glioma dataset regardless of tumor stage was used to build a model, the AUROC in this study reached 0.84.

Mutations in the *BRAF* gene were predictable when a DL model was trained on tissue microarrays (TMAs) in thyroid cancer and tested on WSIs from the independent TCGA thyroid cancer dataset with an AUROC of 0.98, proving the transferability of the method from TMAs to WSIs [89]. When the model was trained and tested within the TCGA thyroid cancer dataset, the AUROC was 0.95 for *BRAF* mutation prediction and 0.88 for *RAS* mutation prediction [88]. Mutations in *BRAF* and *NRAS* oncogenes were also discovered to be distinguishable with DL models in melanoma, with AUROCs of 0.83 and 0.92, respectively [38]. In uveal melanoma, the presence of *BAP1* mutation has achieved performance comparable to that of pathologists, with AUROCs up to 0.99 on the held-out datasets [86,87]. Likewise, *BRAF* mutations are predictable by DL in colorectal cancer, as shown by multiple studies [28,42,81]. In addition, in colorectal cancer, mutations in a panel of clinically relevant mutations in *APC*, *KRAS*, *PIK3CA*, *SMAD4*, and *TP53* were predictable from conventional histology in several studies [28,62,63]. Moreover, pan-cancer studies analyzing dozens of cancer types at once have further supported the potential of DL-based tools in genetic alteration prediction tasks [37,81,82].

We found only single studies published in the field of DL-based mutation prediction in hematological neoplasms and in endometrial, ovarian, and prostate cancers, respectively. Brück *et al* used bone marrow histopathology images to detect important genetic

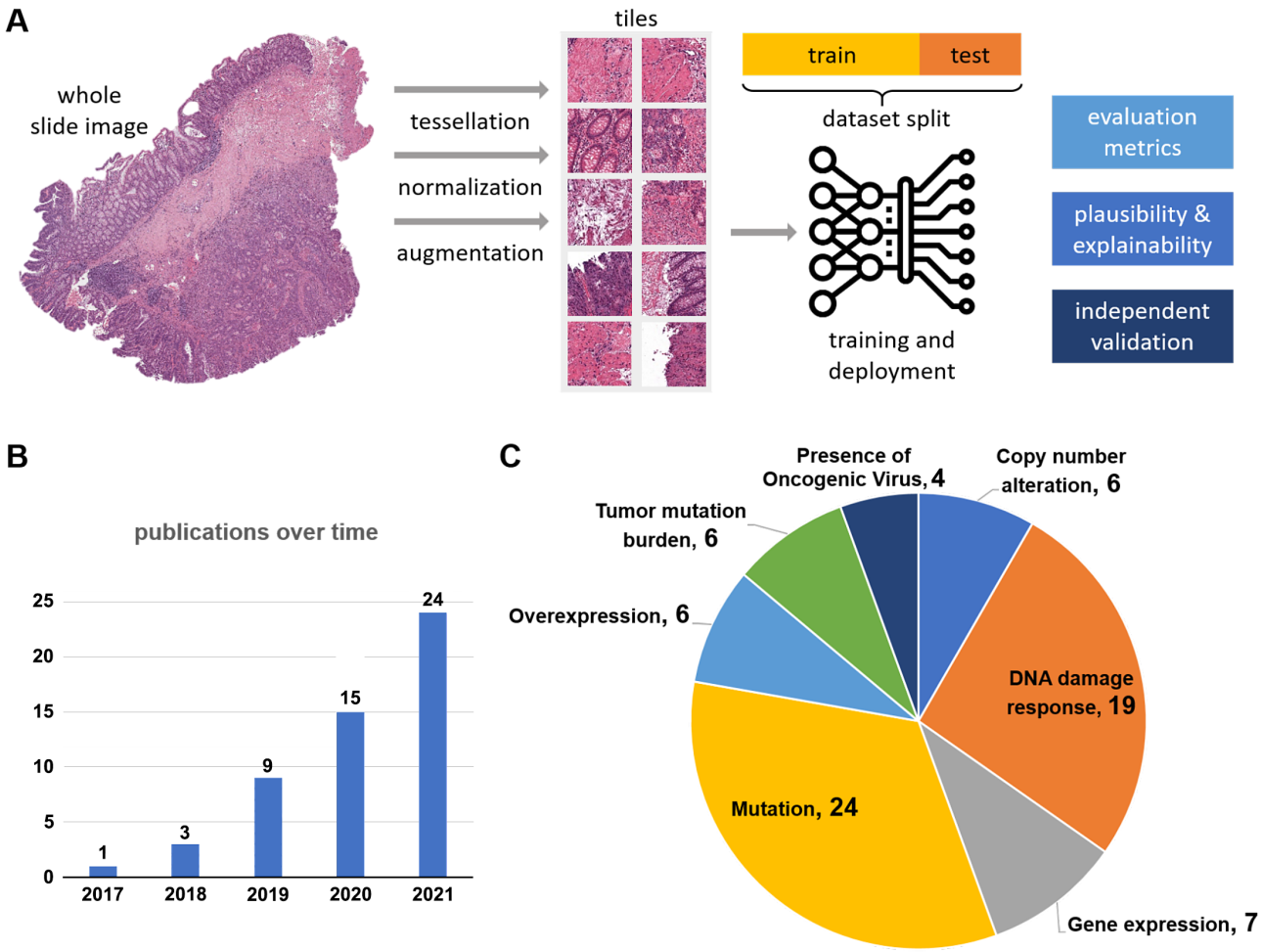

Figure 1. Prediction of molecular alterations from conventional histopathology is a frequently studied task. (A) An overview of deep learning-based AI frameworks in histopathology. (B) Numbers of relevant publications over time. (C) Proportions of genetic alterations reviewed in this study. The numbers next to genetic alteration categories represent the number of publications found by the literature search. Some publications were assigned to multiple categories. Fifty-two unique publications were included in our study. The cutoff date for the quantitative analysis was 28 December 2021. Icon source: Flaticon.com.

features in myelodysplastic syndrome and myeloproliferative neoplasm to predict mutations in genes regulating the cell cycle, cell differentiation, DNA chromatin structure, and *RAS* pathway and mutations in *IDH1*, *IDH2*, *NRAS*, *KRAS*, and spliceosome [50]. In endometrial cancer, polymerase ε (*POLE*) ultra-mutated, MSI-high hypermutated, copy number-low (CNV-L), copy number-high (CNV-H) subtypes, and the mutation status of 18 endometrial carcinoma-related genes were predictable using histopathology images as input [67]. Zeng *et al* developed a method that detects *BRCA1* and *BRCA2* mutations with AUROC values of 0.95 and 0.91, respectively, on the held-out dataset in high-grade serous ovarian carcinoma and also showed that integrating genomics, transcriptomics, and proteomics data with the image features leads to better prognostic models compared with images alone [80]. While prostate cancer is one of the most frequent tumor types in males, computational pathology studies have mainly focused on tumor detection, rather than molecular characterization [97,98]. This could be due to the lack of clinically actionable molecular alterations in this tumor entity. The only study in prostate cancer aimed to predict *SPOP* mutations, which are associated with a better therapeutic response, resulting in an AUROC of 0.86 on an external dataset [85,99].

Together, these findings show that DL can predict a range of clinically relevant mutations directly from H&E slides in multiple tumor types. Work across different tumor types is likely to expand with the emergence of new predictive biomarkers in the future. The current evidence has the principal limitation that the classification performance in almost all cases is still considerably lower than gold standard methods. Nevertheless, by choosing a high-sensitivity, low-specificity operating point, these methods could be used as pre-screening tools for rare mutations.

Tumor mutation burden

Tumor mutation burden (TMB) is defined as the number of somatic mutations accumulated within the tumor cells, and has been associated with prognosis and response to therapy response [100,101]. Thus, TMB is strongly correlated with immunogenicity, due to the

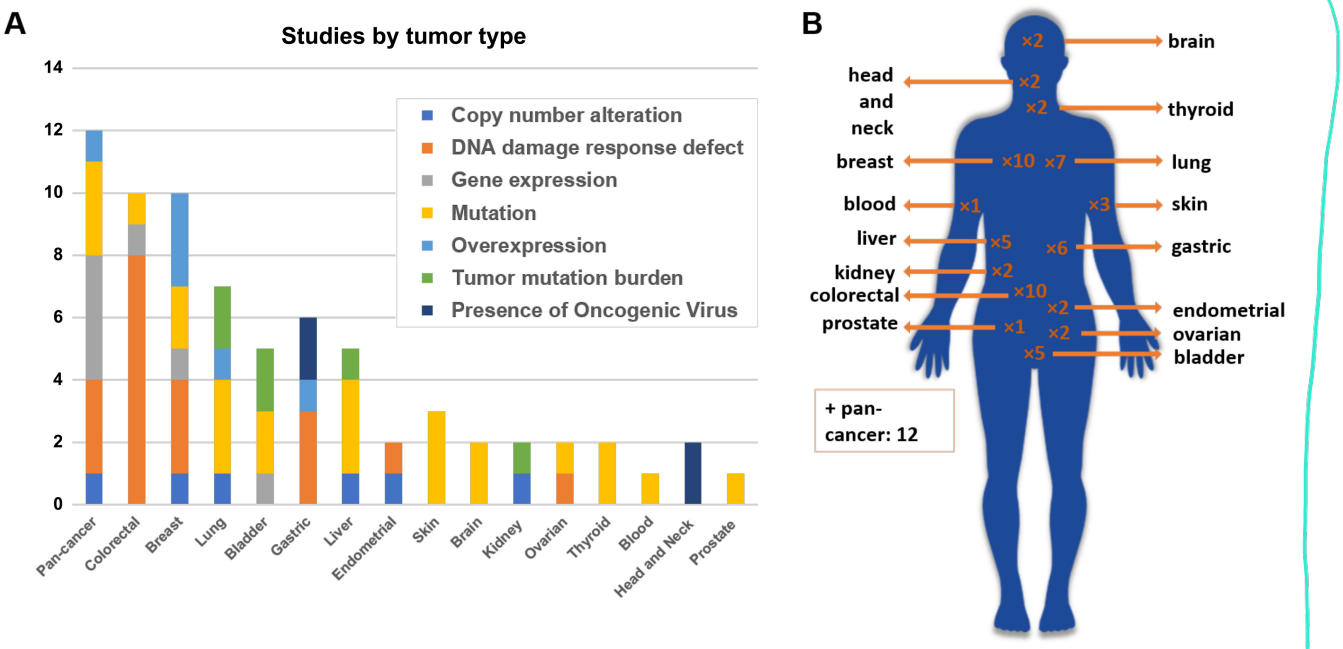

Figure 2. Analysis of primary tumor site in molecular prediction studies. (A) Distribution of the publications based on tumor type. (B) An illustration of the distributions of tumor types in the body. Icon source: Flaticon.com.

increased likelihood of neoantigen presence on the surface of tumor cells with a high TMB [102,103]. Neoantigens are the antigens that are found only on the cell surface of tumor cells produced by somatic mutation; hence, they evoke a T-cell response and are highly immunogenic [104,105]. Although not all mutations produce neoantigens, with the increase in the mutational load, as in the case of cells with a high TMB, the likelihood of the presence of neoantigens increases [102]. Consequently, high TMB is an FDA-approved biomarker for immune checkpoint inhibitors [106]. Owing to the fact that TMB tumors are inflamed, and possibly due to other morphological changes, multiple studies have shown that it is possible to infer TMB from H&E histology [43,45,73,74,79]. The first attempt to predict TMB status from WSIs was made by Zhang *et al* in 2019 [74], where TMB-low and -high groups were targeted on histopathology images from patients with liver cancer. This study reported an AUROC of 0.95 on the validation dataset and showed that TMB prediction by the model outperformed a model based on next-generation sequencing (NGS) [74]. Similarly, TMB status has been shown to be predictable as a binary classification task on urothelial bladder carcinoma, with an AUROC of 0.75 [45], and lung adenocarcinoma, with AUROCs of 0.74 and 0.71 by cross-validated analyses [45,79]. By utilizing predictive TMB scores in survival outcome prediction, Xu *et al* reinforced the prognostic value of TMB [45]. While these studies focused on developing methods for binary classification of TMB status as high and low, Marostica *et al* performed a regression task for TMB prediction on clear cell renal cell carcinoma, meaning that the model is built to predict the TMB score itself, and found a Spearman correlation coefficient of 0.419 on the held-out test dataset [73].

TMB measurement has faced many challenges in clinical decision-making, due to costly NGS-based tests, limited tissue availability, and intratumor TMB variation [45,74]. The publications in the context of TMB prediction, therefore, have great potential for clinical decision-making. All studies reviewed here, however, lack an external test dataset, which is a prerequisite for moving towards clinical application [94].

Defective DNA repair mechanisms: MSI and HRD

The mismatch repair system and the homologous repair system are two biologically and clinically relevant ways of cells to repair DNA damage. Deficiencies in either of these systems can render solid tumors susceptible to specific treatment types.

Microsatellites are the genomic regions consisting of short tandem repeats that are highly susceptible to built-up replication errors in the presence of deficient mismatch repair (dMMR) leading to the MSI phenotype [107]. In some tumor types, such as colorectal, gastric, and endometrial cancer, MSI is the most important determinant of immunogenicity. This is associated with a high TMB, increase in infiltrating lymphocytes, and immunogenic neoantigens expressed due to frameshift mutations found in MSI-high tumors [108–110]. Therefore, high MSI status is approved as a biomarker in many solid tumors for immunotherapy [108,111]. Even before the DL era, pathological predictors of MSI in colorectal cancer were known [12]. Consecutively, it was demonstrated that DL can predict MSI status from histology as well [28,36,57,60,61,64,66,67,69,83,112]. When a DL system is trained on thousands of patients, the predictive power is superb and can reach AUROCs of above 0.95 [61,62], and a recent head-to-head comparison

showed that DL outperforms pathologists [66]. In ovarian cancer, the predictive power of DL models is comparatively much higher, as they reached AUROC values of 0.92 in the held-out test datasets, where the sample size was 114 with the model fed with a training dataset of 115 patients [80]. Newer technologies such as multiple instance learning (MIL) and self-supervised learning (SSL) have improved classification performance by more than 10% compared with the conventional WSI classification model with a sample size as low as 100 patients in colorectal cancer [57]. Schrris *et al* showed that the performance of the MSI classifier can reach above 0.90 in colorectal cancer by combining SSL with an attention-based DL model [57].

Homologous recombination deficiency (HRD), like MSI, is an indicator of a defect of DNA damage repair mechanisms. Homologous recombination is one of the mechanisms within the cell cycle to repair double-stranded DNA breaks, ensuring genomic integrity [113,114]. Tumor cells with HRD show increased base excision repair/single-strand break repair (BER/SSBR) pathways that rely on poly(ADP-ribose) polymerase (PARP) proteins as essential components to cope with DNA damage [115,116]. Identification of HRD in tumors has clinically relevant importance, as therapeutic options targeting PARP activity exist for breast, ovarian, pancreatic, and prostate cancer [114]. The studies that try to predict HRD from WSIs via DL algorithms focused primarily on breast cancer, where the classification of HRD-deficient and HRD-proficient patients was initially predicted with an approximate AUROC of 0.70 [55,84]. Further, Diao *et al* demonstrated that binarized HRD score prediction was possible with AUROCs of 0.77 and 0.68 in breast and gastric cancer, respectively, on held-out datasets by a model using human interpretable features [84]. When SSL-based DL approaches were utilized, the model performance peaked, as was also shown in MSI prediction, and AUROC values over 0.80 were reached [56,57].

In summary, the performance of DL systems to predict MSI status is very high. This could be one of the first biomarkers to enter the diagnostic routine [117]. Nevertheless, regulatory approval of these systems may require additional validation studies.

Gene expression and copy number alterations

All previously discussed methods are *weakly supervised*: The DL classifiers use a class for the entire whole-slide image but process the image in small patches, or tiles. When the classifier is deployed, tile-level predictive scores are aggregated into the patient-level prediction score. These methods essentially lack spatial resolution. To tackle this challenge, He *et al* developed an algorithm where the DL model predicted the expression of over 200 genes for each tile in breast cancers and validated their results on an external dataset [54]. The model in this research, trained with tiles labeled with spatial transcriptomic data, was able to predict breast cancer biomarkers with good accuracy while shedding light on the intratumor heterogeneity in terms of targeted gene expression.

Gene expression prediction on histopathology images has also been possible by research where patient-level expression was integrated into the DL analysis. Schmauch *et al* introduced a framework that fed the DL model with WSI tiles labeled with bulk normalized RNA sequencing data per slide and predicted the slide-level gene expression scores in hepatocellular carcinomas and invasive breast carcinomas [83]. The well-predicted gene expressions by this approach were observed to be involved in the pathways peculiar to both cancers. Similarly, binarized (high versus low) gene expressions quantified by RNA sequencing values per slide were predicted in the study by Diao *et al* [84]. In this study, the model achieved AUROCs greater than 0.65 for prediction of PD-1, PD-L1, CTLA-4, and TIGIT binarized expressions in several different cancer types. Another pan-cancer study by Kather *et al* [81] investigated the prediction of gene expression profiles from conventional histological images, reaching a high performance especially for immune-related gene expression signatures. Similar results were reached by Fu *et al* in another pan-cancer study [37]. Woerl *et al* performed a classification task for molecular subtypes of urothelial bladder carcinoma with a superior accuracy, where the subtypes were determined by gene expression data and assigned to the WSI labels [44]. Similarly, in colorectal cancer, DL is able to predict consensus molecular subtype from conventional images [65]. In all of these bulk expression prediction tasks, spatial expression values are inferred after the analysis is performed.

Copy number variations, which are correlated to gene expression in cancer [118], have also been predicted from H&E images with DL. In a study performed by Qu *et al*, point mutations in *NOTCH1* and *TP53* genes as well as copy number alteration (CNA) status in the *FGFR1* gene in lung adenocarcinoma, and point mutations in *RB1* and *NF1* and *TGFβ2* CNA in liver cancer, were predicted. Of note, the underlying DL model was trained on a breast cancer dataset, showing the partial transferability of the DL models between different organs [58]. The main focus of this study was the prediction of point mutations and CNAs in important genes such as *TP53* and *FGFR*, where the model achieved AUROCs greater than 0.65 for six genes in both categories. *FGFR* mutation status was also shown to be predicted in NSCLC with an AUROC of 0.72 on the held-out dataset [58]. Other studies on external datasets in liver cancer have also shown that *CTNNB1*, *FMN2*, *TP53*, and *ZFX4*, *ALB*, *CSMD3*, *MUC4*, *OBSCN*, and *RYR2* can be detected from WSIs [75,76]. Furthermore, Marostica *et al* used DL methods to predict CNA status in several genes associated with prognosis and with a high prevalence in kidney cancer [73].

In summary, the potential of DL to predict gene expression status and CNA might be clinically useful because gold standard methods can be costly and are not ubiquitously available. Still, as we discuss below, validation studies are needed to determine the clinical utility of these approaches.

Biomarker expression

Immunohistochemistry (IHC) is a histopathological method that is being used to ascertain the amount and distribution of specific antigens within a tissue and is commonly used in diagnosis tasks in cancer [119]. IHC plays an important role in treatment decisions and outcome prediction in breast cancer pathology, as molecular subtypes are determined based on IHC for the biomarkers estrogen receptor (ER), human epidermal growth factor receptor 2 (HER2), and progesterone receptor (PR) [120,121]. Anand *et al* developed a DL algorithm to predict overexpression of HER2 from H&E-stained slides and validated it on an external dataset, achieving an AUROC of 0.76 [53]. HER2 overexpression is associated with a poor prognosis in gastric cancer as well as breast cancer [122,123], and HER2 status was predicted in a within-cohort analysis in a gastric cancer histopathology dataset [68]. Shamai *et al* developed a DL-based method for the prediction of 19 relevant biomarkers, among which ER, HER2, and PR were reported to be at least as precise as IHC [52]. In another study by Couture *et al* [51], the accuracy of ER status prediction was reported as 0.84 in a within-cohort analysis. Similarly, Kather *et al* reported good predictability of ER, PR, and HER2 status in breast cancer from H&E histology as part of a pan-cancer analysis [81]. Overexpression of programmed death-ligand 1 (PD-L1) is another clinically relevant phenomenon, for example in lung cancer [124], and Sha *et al* predicted PD-L1 status with an AUROC of 0.8 in the held-out dataset in NSCLC [77].

Together, these studies demonstrate the potential of AI to predict overexpression of clinically relevant genes directly from H&E histology, which could potentially be used as a rapid pre-screen before IHC is performed.

Presence of oncogenic virus

Oncogenic viruses, i.e. Epstein–Barr virus (EBV), human papillomaviruses (HPVs), hepatitis B and C viruses (HBV and HCV), human T-cell lymphotropic virus-1 (HTLV-1), Kaposi's sarcoma herpesvirus (KSHV), and Merkel cell polyomavirus (MCPyV), cause 15–20% all cancers [125,126]. While they all share the ability to express oncogenic proteins that deteriorate the pathways that lead to cell cycle arrest and apoptosis, they act through their unique and various molecular processes in different tissues [126]. Therefore, viral oncogenesis has a significant impact on diagnosis in the clinic. While the detection of oncogenic virus varies for different species of viruses and tissue types [127], attempts to detect it using DL-based computational methods have been successful. Kather *et al* showed for the first time in 2019 that the presence of EBV in gastric cancer and HPV in head and neck cancer was detectable, with AUROCs of 0.81 and 0.70 on the external test datasets [71]. Klein *et al* also developed a DL algorithm that stratifies patients according to the presence of HPV in oropharyngeal squamous cell carcinoma cancer, with an AUROC of 0.8 in two independent datasets [72]. EBV-associated tumors constitute a distinct molecular subtype of gastric cancer that is associated with a better prognosis [128]. Due to the prevalence and high mortality rate of gastric cancer, EBV detection is imperative in clinical decision-making. EBV detection in gastric cancer via DL approaches was detectable with an AUROC of 0.85 on five external test datasets from Germany and Italy [64]. Furthermore, when EBV prediction was performed on the TCGA gastric cancer cohort, the AUROC was measured as 0.85 for the held-out dataset [70]. Zhang *et al* also demonstrated that cases with higher predictive EBV scores are associated with a better prognosis [70].

These studies show that DL could shed light on the viral etiology of tumors, which is in some cases relevant for patient management.

Perspectives and outlook

Quality control and preprocessing protocols

The data being used in computational pathology, i.e. digitized histopathology images, can be obtained and prepared through a variety of ways, which has to be taken into account during the analysis. Sample preprocessing workflows are important for DL models [129]. Formalin-fixed, paraffin-embedded (FFPE) sections are the most common tissue samples in histopathology, resulting in high-quality images that may take days to be prepared, while frozen sections are the tissue samples that are obtained in urgent situations for a rapid overview of the tumor, providing images within less than an hour, albeit being more prone to the artifacts and disturbances in morphology [130,131]. Most DL studies use FFPE sections [81], but some studies have demonstrated good performance on frozen sections [37]. Also, most studies use surgical resection specimens, which yield large contiguous areas of tumor tissue on a slide [132]. A few studies have compared the performance of DL systems applied to biopsy samples [61,93] or virtual biopsies [69] and found a decline of predictive performance. In the future, as DL classifiers are moving towards implementation in clinical routines, it is essential to specify for each classifier which sample types it was trained on and validated on.

The classical patch-based weakly supervised approach

Histological images are large and have to be tessellated before processing because the size of the WSI is too large [133]. The initial study by Coudray *et al* [29] established a simple yet powerful workflow for the prediction of molecular alterations from such WSIs: the 'patch-based' weakly supervised workflows ('vanilla workflow'). In this approach, histopathological images are cut into tiles. These tiles are processed separately with a convolutional neural network (CNN). Some studies have investigated the best CNN architecture for such

tasks, with residual networks ('resnets') giving a good tradeoff between performance and computational effort [81,134]. Some studies have opted to train a CNN [81], while others have used CNNs to extract 'features' and trained classifiers on these features [37]. The results of these approaches are typically similar [135]. In all approaches, tile-level predictions are ultimately pooled for each slide by some type of averaging [133]. In general, these weakly supervised workflows are efficient because they only require a single ground truth for the whole slide. This saves the time of experts and allows morphological features not yet known to be incorporated into the decision-making process [27,69,136]. Initially, many studies used manual tumor annotations (outlines, masks) before the actual DL process [29,36]. However, subsequent work has shown that this is not strictly necessary; even the classical patch-based approach performs fairly well without any manual annotations [37,69,81].

Transition towards new technologies

While the classical patch-based approach is still used in some recent studies, others have proposed other solutions for the problem of pooling tile-level predictions. In particular, multiple instance learning (MIL) is now commonly utilized in this field, due to its high efficiency in weakly supervised settings [57,97,137]. The MIL framework considers the tiles as instances and predicts target labels on the bags that are made up of instances coming from WSIs of the same patients [138,139]. While the naive MIL approach is sensitive to artifacts and outliers, attention-based MIL is more robust. Attention MIL learns the contribution of individual tiles in the decision-making process [137,140]. Most recently, transformers and graph neural networks that are intrinsically trained with the correlation information between different tiles along with the tile images have been proposed [134,141–143]. Another approach that is becoming more and more common in DL systems in histopathology is contrastive SSL, a subset of unsupervised learning [56,57,144,145]. In contrastive self-supervised training, the model learns the patterns in a dataset in the absence of any labels by contrasting different images and rewarding similar images; therefore, it is aimed to obtain better representations of images [146,147]. SSL-based models can combat the scarcity of labeled data and reduce the required sample size, which weakly supervised methods often suffer from [57]. In summary, the pace of technical innovation in computational pathology is very high and further performance gains due to technical improvements can be expected in the next few years.

Interpretability and explainability

A typical neural network used for image analysis can have up to dozens of millions of parameters, which makes it practically impossible to track the steps of an algorithm or understand why the decision being made was made by human experts. Neural networks are consequently referred to as 'black boxes' [148]. This feature brings about potential risks in DL methods, such as biases in the decision-making process going undetected. Hence, explainability, that is, understanding why the model works the way it does, is often a part of DL studies in computational pathology [147]. Another term in this niche, interpretability, on the other hand, describes the side of human understanding of how an AI model works [149]. More interpretable models are simpler models with a small number of parameters so that people can easily follow the decision making. Since more complex and more opaque models arguably lead to better accuracy, it is justified to state that there is an arms race between interpretability and accuracy [150,151]. While it is debatable if interpretability and explainability are an absolute requirement for AI in medicine, there are two important benefits in histopathology by making AI systems explainable. First, it can serve as a tool that helps clinicians to verify that the model has learned reasonable, previously known, morphological features [151]. Second, new features can be discovered, potentially yielding new mechanistic insight [152–154]. Both aspects can also help to reduce reservations and fears towards AI applications in medicine. Commonly used interpretability methods in digital pathology research are heat maps, class activation maps, t-distributed stochastic neighbor embedding (t-SNE) plots, and generation of top predicted patients and tiles [30,37,155,156]. This has been used to detect previously unknown patterns: For instance, Brockmoeller *et al* showed that the deep learning model trained to detect colorectal cancer tissues with lymph node metastasis based its predictions significantly on inflamed fat tissue [153]. In a study by Loeffler *et al*., the heat maps that are generated to reflect the decision of the DL model on each tile distinguish between a tumor area with an *FGFR3* mutation and the wild type, confirming that the model has learned clinically important features in the decision making [46]. Future studies could therefore use DL to discover novel morphological–molecular associations.

Inference of tumor clonality and spatial heterogeneity

Intratumor heterogeneity is an intrinsic feature of cancers that represents the spatial and temporal variations of cell populations within the same tumor [157]. This variation inside the tumor eventually becomes the driving force for natural selection under the pressure of the cancer therapies and might lead to resistant cells surviving, thereby reducing treatment effectiveness [158,159]. Hence, a better understanding and assessment of intratumor heterogeneity could improve therapeutic approaches. One of the key benefits of DL-based prediction of genetic alterations is that predictions are spatially resolved even if the ground truth data are not. This means that it is possible to map the prediction scores of different tiles onto the WSI. Such an opportunity allows experts, first, to understand the decision process within the machine that is critical for the clinic and the legal perspectives and, second, to make inferences about tumor

heterogeneity. It has been shown that such DL-based spatial heterogeneity reflects an underlying genetic heterogeneity in bladder cancer [46,54]. Specifically, DL has also been used to uncover genetic heterogeneity on a fine-grained level in lung cancer [160] and other tumors [161]. Moreover, the intratumor heterogeneity index based on gene expression scores predicted by DL is used to estimate patient survival outcomes on breast and lung cancers, where the association between intratumor heterogeneity and poor survival has been shown once again [162].

Limitations

The major limitation of most studies is the low performance relative to the gold standard. At the moment, DL prediction of genetic biomarkers can be achieved with an AUROC of up to ~0.8 (Table 1). A few notable examples include MSI status – which is known to have a strong morphological correlation [66] and readily available training datasets [93]. It is possible that with larger cohorts, prediction of other biomarkers will also approach or exceed an AUROC of 0.9, but this will require future efforts to collect such datasets.

A fundamental limitation of AI in medicine is the generally low level of evidence. As Kleppe *et al* pointed out, few studies provide an unbiased performance estimation (level III study according to Kleppe) and an unbiased estimation of medical utility (level IV study) [94]. In particular, lack of external validation is a challenge for DL-based analysis in cancer research. Without external validation, it is not possible to assess the true performance of the developed model as it will never have been tested on data that it had not encountered while training. Moreover, both external and training datasets must be chosen in a way that they represent the target patients as well as possible so that overfitting is prevented while ensuring the generalizability of the model. Confounders are also a relevant problem, especially for studies that overwhelmingly rely on TCGA data [161]. Site-specific signatures such as differences in staining techniques, scanners, and/or the population distribution of the hospital submitting the histological dataset become an important confounder within this research, where some genetic targets predicted using TCGA data of several cancer types became unpredictable when the training was held in a site-preserved manner [163].

Implementation in routine workflows

Unlike radiology, histopathology departments across the world are still largely based on handling glass slides, as opposed to digital images. This will conceivably change in the next few years and indeed a fully digital workflow is a prerequisite for useful integration of DL in daily routine. On the other hand, DL is also a tangible incentive to digitize routine workflows. If DL can pre-screen samples for genetic alterations, and thus reduce the total load of molecular testing, this could yield a quantifiable value that justifies the investment into a digital infrastructure. In addition, if DL can improve patient outcomes (i.e. by improving response prediction), this would also be a strong argument for such investments. Therefore, in the next few years, it will be important to perform rigorous studies aimed at providing strong evidence for DL reaching hard endpoints such as cost savings and patient outcomes. Another requirement is the regulatory approval of DL systems, for which broad validation and explainability of new biomarkers are helpful. If this succeeds, DL will be the pivotal reason for digitalization, which the digital pathology community has been waiting for since the 1990s.

Acknowledgements

JNK is supported by the German Federal Ministry of Health (DEEP LIVER, ZMVI1-2520DAT111) and the Max-Eder-Programme of the German Cancer Aid (grant #70113864). No other specific funding for this work is declared by any of the authors.

Author contributions statement

All the authors designed the study. DC performed the literature search and the analysis. All the authors jointly wrote the manuscript and approved the final version of the manuscript.

References

1. Souza da Silva RM, Queiroga EM, Paz AR, *et al*. Standardized assessment of the tumor–stroma ratio in colorectal cancer: interobserver validation and reproducibility of a potential prognostic factor. *Clin Pathol* 2021; **14:** 2632010X2198968.
2. Galon J, Costes A, Sanchez-Cabo F, *et al*. Type, density, and location of immune cells within human colorectal tumors predict clinical outcome. *Science* 2006; **313:** 1960–1964.
3. Lu Z, Xu S, Shao W, *et al*. Deep-learning-based characterization of tumor-infiltrating lymphocytes in breast cancers from histopathology images and multiomics data. *JCO Clin Cancer Inform* 2020; **4:** 480–490.
4. Shaban M, Raza SEA, Hassan M, *et al*. A digital score of tumour-associated stroma infiltrating lymphocytes predicts survival in head and neck squamous cell carcinoma. *J Pathol* 2022; **256:** 174–185.
5. 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 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:** 235–251.
6. Amgad M, Stovgaard ES, Balslev E, *et al*. Report on computational assessment of tumor infiltrating lymphocytes from the International Immuno-Oncology Biomarker Working Group. *npj Breast Cancer* 2020; **6:** 16.
7. 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:** 1155–1164.
8. Acs B, Salgado R, Hartman J. What do we still need to learn on digitally assessed biomarkers? *eBioMedicine* 2021; **70:** 103520.
9. Sirinukunwattana K, Snead D, Epstein D, *et al*. Novel digital signatures of tissue phenotypes for predicting distant metastasis in colorectal cancer. *Sci Rep* 2018; **8:** 13692.
10. Richards CH, Roxburgh CSD, Anderson JH, *et al*. Prognostic value of tumour necrosis and host inflammatory responses in colorectal cancer. *Br J Surg* 2012; **99:** 287–294.
11. Pai RK, Jayachandran P, Koong AC, *et al*. BRAF-mutated, microsatellite-stable adenocarcinoma of the proximal colon: an aggressive adenocarcinoma with poor survival, mucinous differentiation, and adverse morphologic features. *Am J Surg Pathol* 2012; **36:** 744–752.
12. Greenson JK, Huang S-C, Herron C, *et al*. Pathologic predictors of microsatellite instability in colorectal cancer. *Am J Surg Pathol* 2009; **33:** 126–133.
13. Sofi GN, Sofi JN, Nadeem R, *et al*. Estrogen receptor and progesterone receptor status in breast cancer in relation to age, histological grade, size of lesion and lymph node involvement. *Asian Pac J Cancer Prev* 2012; **13:** 5047–5052.
14. Ninomiya H, Hiramatsu M, Inamura K, *et al*. Correlation between morphology and EGFR mutations in lung adenocarcinomas: significance of the micropapillary pattern and the hobnail cell type. *Lung Cancer* 2009; **63:** 235–240.
15. Metter DM, Colgan TJ, Leung ST, *et al*. Trends in the US and Canadian pathologist workforces from 2007 to 2017. *JAMA Netw Open* 2019; **2:** e194337.
16. Hamilton PW, Bankhead P, Wang Y, *et al*. Digital pathology and image analysis in tissue biomarker research. *Methods* 2014; **70:** 59–73.
17. Tuominen VJ, Ruotoistenmäki S, Viitanen A, *et al*. ImmunoRatio: a publicly available web application for quantitative image analysis of estrogen receptor (ER), progesterone receptor (PR), and Ki-67. *Breast Cancer Res* 2010; **12:** R56.
18. Codella N, Moradi M, Matasar M, *et al*. Lymphoma diagnosis in histopathology using a multi-stage visual learning approach. *Proc SPIE Volume 9791. Medical Imaging 2016: Digital Pathology*, 2016; 131–137.
19. Aswathy MA, Jagannath M. An SVM approach towards breast cancer classification from H&E-stained histopathology images based on integrated features. *Med Biol Eng Comput* 2021; **59:** 1773–1783.
20. Leslie LS, Wrobel TP, Mayerich D, *et al*. High definition infrared spectroscopic imaging for lymph node histopathology. *PLoS One* 2015; **10:** e0127238.
21. Ren J, Sadimin E, Foran DJ, *et al*. Computer aided analysis of prostate histopathology images to support a refined Gleason grading system. *Proc SPIE Int Soc Opt Eng* 2017; **10133:** 101331V.
22. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. *Adv Neural Inf Process Syst* 2012; **25:** 1097–1105.
23. Esteva A, Kuprel B, Novoa RA, *et al*. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 2017; **542:** 115–118.
24. Silver D, Schrittwieser J, Simonyan K, *et al*. Mastering the game of Go without human knowledge. *Nature* 2017; **550:** 354–359.
25. Vinyals O, Babuschkin I, Czarnecki WM, *et al*. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature* 2019; **575:** 350–354.
26. Skrede O-J, De Raedt S, Kleppe A, *et al*. Deep learning for prediction of colorectal cancer outcome: a discovery and validation study. *Lancet* 2020; **395:** 350–360.
27. Wulczyn E, Steiner DF, Xu Z, *et al*. Deep learning-based survival prediction for multiple cancer types using histopathology images. *PLoS One* 2020; **15:** e0233678.
28. Schrammen PL, Ghaffari Laleh N, Echle A, *et al*. Weakly supervised annotation-free cancer detection and prediction of genotype in routine histopathology. *J Pathol* 2022; **256:** 50–60.
29. 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:** 1559–1567.
30. Foersch S, Eckstein M, Wagner D-C, *et al*. Deep learning for diagnosis and survival prediction in soft tissue sarcoma. *Ann Oncol* 2021; **32:** 1178–1187.
31. Lucas M, Jansen I, van Leeuwen TG, *et al*. Deep learning-based recurrence prediction in patients with non-muscle-invasive bladder cancer. *Eur Urol Focus* 2022; **8:** 165–172.
32. 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–E2979.
33. Courtiol P, Maussion C, Moarii M, *et al*. Deep learning-based classification of mesothelioma improves prediction of patient outcome. *Nat Med* 2019; **25:** 1519–1525.
34. Saillard C, Schmauch B, Laifa O, *et al*. Predicting survival after hepatocellular carcinoma resection using deep learning on histological slides. *Hepatology* 2020; **72:** 2000–2013.
35. Nam D, Chapiro J, Paradis V, *et al*. Artificial intelligence in liver diseases: improving diagnostics, prognostics and response prediction. *JHEP Rep* 2022; **4:** 100443.
36. Kather JN, Pearson AT, Halama N, *et al*. Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. *Nat Med* 2019; **25:** 1054–1056.
37. Fu Y, Jung AW, Torne RV, *et al*. Pan-cancer computational histopathology reveals mutations, tumor composition and prognosis. *Nat Cancer* 2020; **1:** 800–810.
38. Kim RH, Nomikou S, Dawood Z, *et al*. A deep learning approach for rapid mutational screening in melanoma. *bioRxiv* 2020. doi.org/10.1101/610311. [Not peer reviewed].
39. Taylor-Weiner A, Pedawi A, Chui WF, *et al*. Abstract PD6-04: deep-learning based prediction of homologous recombination deficiency (hrd) status from histological features in breast cancer; a research study. *Cancer Res* 2021; **81:** PD6-04.
40. Galateau Salle F, Le Stang N, Tirode 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* 2020; **15:** 1037–1053.
41. Arslan S, Mehrotra D, Schmidt J, *et al*. Large-scale systematic feasibility study on the pan-cancer predictability of multi-omic biomarkers from whole slide images with deep learning. *bioRxiv* 2022. doi.org/10.1101/2022.01.21.477189. [Not peer reviewed].
42. Bilal M, Raza SEA, Azam A, *et al*. Development and validation of a weakly supervised deep learning framework to predict the status of molecular pathways and key mutations in colorectal cancer from routine histology images: a retrospective study. *Lancet Digit Health* 2021; **3:** e763–e772.
43. Park S, Xu H, Hwang TH. Gaussian process based heteroscedastic noise modeling for tumor mutation burden prediction from whole slide images. *bioRxiv* 2019. doi.org/10.1101/554261. [Not peer reviewed].
44. Woerl A-C, Eckstein M, Geiger J, *et al*. Deep learning predicts molecular subtype of muscle-invasive bladder cancer from conventional histopathological slides. *Eur Urol* 2020; **78:** 256–264.
45. Xu H, Park S, Clemenceau JR, *et al*. Spatial heterogeneity and organization of tumor mutation burden and immune infiltrates within tumors based on whole slide images correlated with patient survival in bladder cancer. *bioRxiv* 2020. doi.org/10.1101/554527. [Not peer reviewed].
46. Loeffler CML, Ortiz Bruechle N, Jung M, *et al*. Artificial intelligence-based detection of FGFR3 mutational status directly

from routine histology in bladder cancer: a possible preselection for molecular testing? *Eur Urol Focus* 2021; **8**: 472–479.
47. Velmahos CS, Badgeley M, Lo Y-C. Using deep learning to identify bladder cancers with *FGFR*-activating mutations from histology images. *Cancer Med* 2021; **10:** 4805–4813.
48. Cui D, Liu Y, Liu G, *et al*. A multiple-instance learning-based convolutional neural network model to detect the mutation in the histopathology images of glioma tissues. *J Comput Biol* 2020; **27:** 1264–1272.
49. Jiang S, Zanazzi GJ, Hassanpour S. Predicting prognosis and IDH mutation status for patients with lower-grade gliomas using whole slide images. *Sci Rep* 2021; **11:** 16849.
50. Brück OE, Lallukka-Brück SE, Hohtari HR, *et al*. Machine learning of bone marrow histopathology identifies genetic and clinical determinants in patients with MDS. *Blood Cancer Discov* 2021; **2:** 238–249.
51. Couture HD, Williams LA, Geradts J, *et al*. Image analysis with deep learning to predict breast cancer grade, ER status, histologic subtype, and intrinsic subtype. *NPJ Breast Cancer* 2018; **4:** 30.
52. Shamai G, Binenbaum Y, Slossberg R, *et al*. Artificial intelligence algorithms to assess hormonal status from tissue microarrays in patients with breast cancer. *JAMA Netw Open* 2019; **2:** e197700.
53. Anand D, Kurian NC, Dhage S, *et al*. Deep learning to estimate human epidermal growth factor receptor 2 status from hematoxylin and eosin-stained breast tissue images. *J Pathol Inform* 2020; **11:** 19.
54. He B, Bergenstråhle L, Stenbeck L, *et al*. Integrating spatial gene expression and breast tumour morphology via deep learning. *Nat Biomed Eng* 2020; **4:** 827–834.
55. Valieris R, Amaro L, CAB de T O, *et al*. Deep learning predicts underlying features on pathology images with therapeutic relevance for breast and gastric cancer. *Cancers (Basel)* 2020; **12:** 3687.
56. Lazard T, Bataillon G, Naylor P, *et al*. Deep learning identifies new morphological patterns of Homologous Recombination Deficiency in luminal breast cancers from whole slide images. *bioRxiv* 2021; doi.org/10.1101/2021.09.10.459734. [Not peer reviewed].
57. Schirris Y, Gavves E, Nederlof I, *et al*. DeepSMILE: Self-supervised heterogeneity-aware multiple instance learning for DNA damage response defect classification directly from H&E whole-slide images. *arXiv* 2021; arXiv:2107.09405[eessIV]. [Not peer reviewed].
58. Qu H, Zhou M, Yan Z, *et al*. Genetic mutation and biological pathway prediction based on whole slide images in breast carcinoma using deep learning. *NPJ Precis Oncol* 2021; **5:** 87.
59. Wang X, Zou C, Zhang Y, *et al*. Prediction of BRCA gene mutation in breast cancer based on deep learning and histopathology images. *Front Genet* 2021; **12:** 661109.
60. Cao R, Yang F, Ma S-C, *et al*. Development and interpretation of a pathomics-based model for the prediction of microsatellite instability in colorectal cancer. *Theranostics* 2020; **10:** 11080–11091.
61. Echle A, Grabsch HI, Quirke P, *et al*. Clinical-grade detection of microsatellite instability in colorectal tumors by deep learning. *Gastroenterology* 2020; **159:** 1406–1416.
62. Bilal M, Ahmed Raza SE, Azam A, *et al*. Novel deep learning algorithm predicts the status of molecular pathways and key mutations in colorectal cancer from routine histology images. *medRxiv* 2021. doi.org/10.1101/2021.01.19.21250122. [Not peer reviewed].
63. Jang H-J, Lee A, Kang J, *et al*. Prediction of clinically actionable genetic alterations from colorectal cancer histopathology images using deep learning. *World J Gastroenterol* 2020; **26:** 6207–6223.
64. Lee SH, Song IH, Jang H-J. Feasibility of deep learning-based fully automated classification of microsatellite instability in tissue slides of colorectal cancer. *Int J Cancer* 2021; **149:** 728–740.
65. Sirinukunwattana K, Domingo E, Richman SD, *et al*. Image-based consensus molecular subtype (imCMS) classification of colorectal cancer using deep learning. *Gut* 2021; **70:** 544–554.
66. Yamashita R, Long J, Longacre T, *et al*. Deep learning model for the prediction of microsatellite instability in colorectal cancer: a diagnostic study. *Lancet Oncol* 2021; **22:** 132–141.
67. Hong R, Liu W, DeLair D, *et al*. Predicting endometrial cancer subtypes and molecular features from histopathology images using multi-resolution deep learning models. *Cell Rep Med* 2021; **2:** 100400.
68. Sharma H, Zerbe N, Klempert I, *et al*. Deep convolutional neural networks for automatic classification of gastric carcinoma using whole slide images in digital histopathology. *Comput Med Imaging Graph* 2017; **61:** 2–13.
69. Muti HS, Heij LR, Keller G, *et al*. Development and validation of deep learning classifiers to detect Epstein–Barr virus and microsatellite instability status in gastric cancer: a retrospective multicentre cohort study. *Lancet Digit Health* 2021; **3:** e654–e664.
70. Zhang B, Yao K, Xu M, *et al*. Deep learning predicts EBV status in gastric cancer based on spatial patterns of lymphocyte infiltration. *Cancers (Basel)* 2021; **13:** 6002.
71. Kather JN, Schulte J, Grabsch HI, *et al*. Deep learning detects virus presence in cancer histology. *bioRxiv* 2019. doi.org/10.1101/690206. [Not peer reviewed].
72. Klein S, Quaas A, Quantius J, *et al*. Deep learning predicts HPV association in oropharyngeal squamous cell carcinomas and identifies patients with a favorable prognosis using regular H&E stains. *Clin Cancer Res* 2021; **27:** 1131–1138.
73. Marostica E, Barber R, Denize T, *et al*. Development of a histopathology informatics pipeline for classification and prediction of clinical outcomes in subtypes of renal cell carcinoma. *Clin Cancer Res* 2021; **27:** 2868–2878.
74. Zhang H, Ren F, Wang Z et al. Predicting tumor mutational burden from liver cancer pathological images using convolutional neural network. 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2019; 920–925. doi.org/10.1109/BIBM47256.2019.8983139.
75. Chen M, Zhang B, Topatana W, *et al*. Classification and mutation prediction based on histopathology H&E images in liver cancer using deep learning. *NPJ Precis Oncol* 2020; **4:** 14.
76. Liao H, Long Y, Han R, *et al*. Deep learning-based classification and mutation prediction from histopathological images of hepatocellular carcinoma. *Clin Transl Med* 2020; **10:** e102.
77. Sha L, Osinski BL, Ho IY, *et al*. Multi-field-of-view deep learning model predicts nonsmall cell lung cancer programmed death-ligand 1 status from whole-slide hematoxylin and eosin images. *J Pathol Inform* 2019; **10:** 24.
78. Wang Q, Shen Q, Zhang Z, *et al*. [Prediction of gene mutation in lung cancer based on deep learning and histomorphology analysis]. *Sheng Wu Yi Xue Gong Cheng Xue Za Zhi* 2020; **37:** 10–18 [Article in Chinese].
79. Sadhwani A, Chang H-W, Behrooz A, *et al*. Comparative analysis of machine learning approaches to classify tumor mutation burden in lung adenocarcinoma using histopathology images. *Sci Rep* 2021; **11:** 16605.
80. Zeng H, Chen L, Zhang M, *et al*. Integration of histopathological images and multi-dimensional omics analyses predicts molecular features and prognosis in high-grade serous ovarian cancer. *Gynecol Oncol* 2021; **163:** 171–180.
81. Kather JN, Heij LR, Grabsch HI, *et al*. Pan-cancer image-based detection of clinically actionable genetic alterations. *Nat Cancer* 2020; **1:** 789–799.
82. Noorbakhsh J, Farahmand S, Foroughi Pour A, *et al*. Deep learning-based cross-classifications reveal conserved spatial behaviors within tumor histological images. *Nat Commun* 2020; **11:** 6367.
83. Schmauch B, Romagnoni A, Pronier E, *et al*. A deep learning model to predict RNA-Seq expression of tumours from whole slide images. *Nat Commun* 2020; **11:** 3877.
84. Diao JA, Wang JK, Chui WF, *et al*. Human-interpretable image features derived from densely mapped cancer pathology slides predict diverse molecular phenotypes. *Nat Commun* 2021; **12:** 1613.

85. Schaumberg AJ, Rubin MA, Fuchs TJ. H&E-stained whole slide image deep learning predicts SPOP mutation state in prostate cancer. *bioRxiv* 2018. doi.org/10.1101/064279. [Not peer reviewed].
86. Sun M, Zhou W, Qi X, *et al*. Prediction of BAP1 expression in uveal melanoma using densely-connected deep classification networks. *Cancers (Basel)* 2019; **11:** 1579.
87. Mehdiratta G, Sahai S. Prediction of BAP1 mutations in uveal melanoma patients from histology images using weakly supervised deep learning-based whole slide image analysis. *medRxiv* 2021. doi.org/10.1101/2021.09.16.21263694. [Not peer reviewed].
88. Tsou P, Wu C-J. Mapping driver mutations to histopathological subtypes in papillary thyroid carcinoma: applying a deep convolutional neural network. *J Clin Med Res* 2019; **8:** 1675.
89. Anand D, Yashashwi K, Kumar N, *et al*. Weakly supervised learning on unannotated H&E-stained slides predicts *BRAF* mutation in thyroid cancer with high accuracy. *J Pathol* 2021; **255:** 232–242.
90. Florkowski CM. Sensitivity, specificity, receiver-operating characteristic (ROC) curves and likelihood ratios: communicating the performance of diagnostic tests. *Clin Biochem Rev* 2008; **29**((Suppl 1)): S83–S87.
91. Obuchowski NA. ROC analysis. *Am J Roentgenol* 2005; **184:** 364–372.
92. Halligan S, Altman DG, Mallett S. Disadvantages of using the area under the receiver operating characteristic curve to assess imaging tests: a discussion and proposal for an alternative approach. *Eur Radiol* 2015; **25:** 932–939.
93. Echle A, Ghaffari Laleh N, Quirke P, *et al*. Artificial intelligence for detection of microsatellite instability in colorectal cancer – a multicentric analysis of a pre-screening tool for clinical application. *ESMO Open* 2022; **7:** 100400.
94. Kleppe A, Skrede O-J, De Raedt S, *et al*. Designing deep learning studies in cancer diagnostics. *Nat Rev Cancer* 2021; **21:** 199–211.
95. Walsh MF, Nathanson KL, Couch FJ, *et al*. Genomic biomarkers for breast cancer risk. *Adv Exp Med Biol* 2016; **882:** 1–32.
96. Han S, Liu Y, Cai SJ, *et al*. IDH mutation in glioma: molecular mechanisms and potential therapeutic targets. *Br J Cancer* 2020; **122:** 1580–1589.
97. 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:** 1301–1309.
98. Tolkach Y, Dohmgörgen T, Toma M, *et al*. High-accuracy prostate cancer pathology using deep learning. *Nat Mach Intell* 2020; **2:** 411–418.
99. Clark A, Burleson M. SPOP and cancer: a systematic review. *Am J Cancer Res* 2020; **10:** 704–726.
100. Nagahashi M, Sato S, Yuza K, *et al*. Common driver mutations and smoking history affect tumor mutation burden in lung adenocarcinoma. *J Surg Res* 2018; **230:** 181–185.
101. Wang X, Li M. Correlate tumor mutation burden with immune signatures in human cancers. *BMC Immunol* 2019; **20:** 4.
102. Chan TA, Yarchoan M, Jaffee E, *et al*. Development of tumor mutation burden as an immunotherapy biomarker: utility for the oncology clinic. *Ann Oncol* 2019; **30:** 44–56.
103. Galuppini F, Dal Pozzo CA, Deckert J, *et al*. Tumor mutation burden: from comprehensive mutational screening to the clinic. *Cancer Cell Int* 2019; **19:** 209.
104. Garcia-Garijo A, Fajardo CA, Gros A. Determinants for neoantigen identification. *Front Immunol* 2019; **10:** 1392.
105. Peng M, Mo Y, Wang Y, *et al*. Neoantigen vaccine: an emerging tumor immunotherapy. *Mol Cancer* 2019; **18:** 128.
106. Center for Drug Evaluation, Research. FDA approves pembrolizumab for adults and children with TMB-H solid tumors. [Accessed 23 November 2021]. Available from: https://www.fda.gov/drugs/drug-approvals-and-databases/fda-approves-pembrolizumab-adults-and-children-tmb-h-solid-tumors
107. Evrard C, Tachon G, Randrian V, *et al*. Microsatellite instability: diagnosis, heterogeneity, discordance, and clinical impact in colorectal cancer. *Cancers (Basel)* 2019; **11:** 1567.
108. Chang L, Chang M, Chang HM, *et al*. Microsatellite instability: a predictive biomarker for cancer immunotherapy. *Appl Immunohistochem Mol Morphol* 2018; **26:** e15–e21.
109. Nebot-Bral L, Coutzac C, Kannouche PL, *et al*. Why is immunotherapy effective (or not) in patients with MSI/MMRD tumors? *Bull Cancer* 2019; **106:** 105–113.
110. Roudko V, Cimen Bozkus C, Orfanelli T, *et al*. Shared immunogenic poly-epitope frameshift mutations in microsatellite unstable tumors. *Cell* 2020; **183:** 1634–1649.
111. Yamashita H, Nakayama K, Ishikawa M, *et al*. Microsatellite instability is a biomarker for immune checkpoint inhibitors in endometrial cancer. *Oncotarget* 2017; **9:** 5652–5664.
112. Echle A, Laleh NG, Schrammen PL, *et al*. Deep learning for the detection of microsatellite instability from histology images in colorectal cancer: a systematic literature review. *ImmunoInformatics* 2021; **3–4:** 100008.
113. da Cunha Colombo Bonadio RR, Fogace RN, Miranda VC, *et al*. Homologous recombination deficiency in ovarian cancer: a review of its epidemiology and management. *Clinics (Sao Paulo)* 2018; **73:** e450s.
114. Hoppe MM, Sundar R, Tan DSP, *et al*. Biomarkers for homologous recombination deficiency in cancer. *J Natl Cancer Inst* 2018; **110:** 704–713.
115. Beck C, Robert I, Reina-San-Martin B, *et al*. Poly(ADP-ribose) polymerases in double-strand break repair: focus on PARP1, PARP2 and PARP3. *Exp Cell Res* 2014; **329:** 18–25.
116. Keung M, Wu Y, Vadgama J. PARP inhibitors as a therapeutic agent for homologous recombination deficiency in breast cancers. *J Clin Med* 2019; **8:** 435.
117. Kacew AJ, Strohbehn GW, Saulsberry L, *et al*. Artificial intelligence can cut costs while maintaining accuracy in colorectal cancer genotyping. *Front Oncol* 2021; **11:** 630953.
118. Shao X, Lv N, Liao J, *et al*. Copy number variation is highly correlated with differential gene expression: a pan-cancer study. *BMC Med Genet* 2019; **20:** 175.
119. Duraiyan J, Govindarajan R, Kaliyappan K, *et al*. Applications of immunohistochemistry. *J Pharm Bioallied Sci* 2012; **4:** S307–S309.
120. Blows FM, Driver KE, Schmidt MK, *et al*. Subtyping of breast cancer by immunohistochemistry to investigate a relationship between subtype and short and long term survival: a collaborative analysis of data for 10,159 cases from 12 studies. *PLoS Med* 2010; **7:** e1000279.
121. Zaha DC. Significance of immunohistochemistry in breast cancer. *World J Clin Oncol* 2014; **5:** 382–392.
122. Gravalos C, Jimeno A. HER2 in gastric cancer: a new prognostic factor and a novel therapeutic target. *Ann Oncol* 2008; **19:** 1523–1529.
123. Aman NA, Doukoure B, Koffi KD, *et al*. HER2 overexpression and correlation with other significant clinicopathologic parameters in Ivorian breast cancer women. *BMC Clin Pathol* 2019; **19:** 1.
124. Yu W, Hua Y, Qiu H, *et al*. PD-L1 promotes tumor growth and progression by activating WIP and β-catenin signaling pathways and predicts poor prognosis in lung cancer. *Cell Death Dis* 2020; **11:** 506.
125. Mesri EA, Feitelson MA, Munger K. Human viral oncogenesis: a cancer hallmarks analysis. *Cell Host Microbe* 2014; **15:** 266–282.
126. Krump NA, You J. Molecular mechanisms of viral oncogenesis in humans. *Nat Rev Microbiol* 2018; **16:** 684–698.
127. Sarid R, Gao S-J. Viruses and human cancer: from detection to causality. *Cancer Lett* 2011; **305:** 218–227.
128. Yang J, Liu Z, Zeng B, *et al*. Epstein–Barr virus-associated gastric cancer: a distinct subtype. *Cancer Lett* 2020; **495:** 191–199.

129. Schömig-Markiefka B, Pryalukhin A, Hulla W, *et al*. Quality control stress test for deep learning-based diagnostic model in digital pathology. *Mod Pathol* 2021; **34:** 2098–2108.
130. Sun L, Marsh JN, Matlock MK, *et al*. Deep learning quantification of percent steatosis in donor liver biopsy frozen sections. *eBioMedicine* 2020; **60:** 103029.
131. Kang L, Li X, Zhang Y, *et al*. Deep learning enables ultraviolet photoacoustic microscopy based histological imaging with near real-time virtual staining. *Photoacoustics* 2022; **25:** 100308.
132. Echle A, Rindtorff NT, Brinker TJ, *et al*. Deep learning in cancer pathology: a new generation of clinical biomarkers. *Br J Cancer* 2021; **124:** 686–696.
133. van Treeck M, Cifci D, Laleh NG, *et al*. DeepMed: a unified, modular pipeline for end-to-end deep learning in computational pathology. *bioRxiv* 2021. doi.org/10.1101/2021.12.19.473344. [Not peer reviewed].
134. Laleh NG, Muti HS, Loeffler CML, *et al*. Benchmarking artificial intelligence methods for end-to-end computational pathology. *bioRxiv* 2021. doi.org/10.1101/2021.08.09.455633. [Not peer reviewed].
135. Coudray N, Tsirigos A. Deep learning links histology, molecular signatures and prognosis in cancer. *Nat Cancer* 2020; **1:** 755–757.
136. Courtiol P, Tramel EW, Sanselme M, *et al*. Classification and disease localization in histopathology using only global labels: a weakly-supervised approach. *arXiv* 2018; arXiv:1802.02212 [cs.CV]. [Not peer reviewed].
137. Lu MY, Williamson DFK, Chen TY, *et al*. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nat Biomed Eng* 2021; **5:** 555–570.
138. Ilse M, Tomczak JM, Welling M. Deep multiple instance learning for digital histopathology. In *Handbook of Medical Image Computing and Computer Assisted Intervention*, Zhou SK, Rueckert D, Fichtinger G (eds). Academic Press, 2020; 521–546.
139. Sudharshan PJ, Petitjean C, Spanhol F, *et al*. Multiple instance learning for histopathological breast cancer image classification. *Expert Syst Appl* 2019; **117:** 103–111.
140. Ilse M, Tomczak J, Welling M. Attention-based Deep Multiple Instance Learning. In *Proceedings of the 35th International Conference on Machine Learning*, Volume 80, Dy J, Krause A (eds). Proceedings of Machine Learning Research (PMLR), 2018; 2127–2136.
141. Chen H, Li C, Wang G, *et al*. GasHis-transformer: a multi-scale visual transformer approach for gastric histopathology image classification. *arXiv* 2021; arXiv:2104.14528 [cs.CV]. [Not peer reviewed].
142. Shao Z, Bian H, Chen Y, *et al*. TransMIL: transformer based correlated multiple instance learning for whole slide image classification. *arXiv* 2021; arXiv:2106.00908 [csCV]. [Not peer reviewed].
143. Zheng Y, Gindra R, Betke M, *et al*. A deep learning based graph-transformer for whole slide image classification. *bioRxiv* 2021. doi.org/10.1101/2021.10.15.21265060. [Not peer reviewed].
144. Ciga O, Xu T, Martel AL. Self supervised contrastive learning for digital histopathology. *arXiv* 2020; arXiv:2011.13971 [eessIV]. [Not peer reviewed].
145. Stacke K, Unger J, Lundström C, *et al*. Learning representations with contrastive self-supervised learning for histopathology applications. *arXiv* 2021; arXiv:2112.05760 [eessIV]. [Not peer reviewed].
146. He K, Fan H, Wu Y, *et al*. Momentum contrast for unsupervised visual representation learning. *arXiv* 2019; arXiv:1911.05722 [csCV]. [Not peer reviewed].
147. Chen T, Kornblith S, Norouzi M, *et al*. A simple framework for contrastive learning of visual representations. *arXiv* 2020; arXiv: 2002.05709 [csLG]. [Not peer reviewed].
148. Gaur M, Faldu K, Sheth A, *et al*. Semantics of the black-box: can knowledge graphs help make deep learning systems more interpretable and explainable? *IEEE Internet Computing* 2021; **25:** 51–59.
149. Marcinkevičs R, Vogt JE. Interpretability and explainability: a machine learning zoo mini-tour. *arXiv* 2020; arXiv:2012.01805 [cs.LG]. [Not peer reviewed].
150. Holzinger A, Langs G, Denk H, *et al*. Causability and explainability of artificial intelligence in medicine. *Wiley Interdiscip Rev Data Min Knowl Discov* 2019; **9:** e1312.
151. Escalante HJ, Escalera S, Guyon I, *et al*. (eds). *Explainable and Interpretable Models in Computer Vision and Machine Learning*. Springer: Cham, 2018.
152. Wulczyn E, Steiner DF, Moran M, *et al*. Interpretable survival prediction for colorectal cancer using deep learning. *NPJ Digit Med* 2021; **4:** 71.
153. Brockmoeller S, Echle A, Laleh NG, *et al*. Deep learning identifies inflamed fat as a risk factor for lymph node metastasis in early colorectal cancer. *J Pathol* 2022; **256:** 269–281.
154. Kim RH, Nomikou S, Coudray N, *et al*. Deep learning and pathomics analyses reveal cell nuclei as important features for mutation prediction of BRAF-mutated melanomas. *J Invest Dermatol* 2022; **142:** 1650–1658.e6.
155. Gamble P, Jaroensri R, Wang H, *et al*. Determining breast cancer biomarker status and associated morphological features using deep learning. *Commun Med* 2021; **1:** 14.
156. Naik N, Madani A, Esteva A, *et al*. Deep learning-enabled breast cancer hormonal receptor status determination from base-level H&E stains. *Nat Commun* 2020; **11:** 5727.
157. Dagogo-Jack I, Shaw AT. Tumour heterogeneity and resistance to cancer therapies. *Nat Rev Clin Oncol* 2017; **15:** 81–94.
158. Andor N, Graham TA, Jansen M, *et al*. Pan-cancer analysis of the extent and consequences of intratumor heterogeneity. *Nat Med* 2016; **22:** 105–113.
159. Marusyk A, Janiszewska M, Polyak K. Intratumor heterogeneity: the Rosetta stone of therapy resistance. *Cancer Cell* 2020; **37:** 471–484.
160. AbdulJabbar K, Raza SEA, Rosenthal R, *et al*. Geospatial immune variability illuminates differential evolution of lung adenocarcinoma. *Nat Med* 2020; **26:** 1054–1062.
161. Zormpas-Petridis K, Noguera R, Ivankovic DK, *et al*. SuperHistopath: a deep learning pipeline for mapping tumor heterogeneity on low-resolution whole-slide digital histopathology images. *Front Oncol* 2020; **10:** 586292.
162. Levy-Jurgenson A, Tekpli X, Kristensen VN, *et al*. Spatial transcriptomics inferred from pathology whole-slide images links tumor heterogeneity to survival in breast and lung cancer. *Sci Rep* 2020; **10:** 18802.
163. Howard FM, Dolezal J, Kochanny S, *et al*. The impact of site-specific digital histology signatures on deep learning model accuracy and bias. *Nat Commun* 2021; **12:** 4423.
164. Ouzzani M, Hammady H, Fedorowicz Z, *et al*. Rayyan – a web and mobile app for systematic reviews. *Syst Rev* 2016; **5:** 1–10.

Reference 164 is cited only in the supplementary material.

SUPPLEMENTARY MATERIAL ONLINE

Supplementary methods

Figure S1. Overview of literature review