Seminars in Cancer Biology 97 (2023) 70–85

Contents lists available at ScienceDirect

Seminars in Cancer Biology

journal homepage: www.elsevier.com/locate/semcancer

One label is all you need: Interpretable AI-enhanced histopathology for oncology

Thomas E. Tavolara, Ziyu Su, Metin N. Gurcan, M. Khalid Khan Niazi *

Center for Artificial Intelligence Research, Wake Forest University School of Medicine, Winston-Salem, NC, USA

ARTICLE INFO

Keywords:
Artificial intelligence
Weakly supervised
Multiple instance learning
Histopathology
Verifiable AI
Interpretable AI
Explainable AI
Morphological markers
Tumor detection
Metastasis
Tumor subtyping
Tumor grading
Molecular markers
Gene expression
Molecular subtyping
Genetic markers
Mutational burden
Microsatellite instability
Chromosomal instability
Homologous recombination deficiency

ABSTRACT

Artificial Intelligence (AI)-enhanced histopathology presents unprecedented opportunities to benefit oncology through interpretable methods that require only one overall label per hematoxylin and eosin (H&E) slide with no tissue-level annotations. We present a structured review of these methods organized by their degree of verifiability and by commonly recurring application areas in oncological characterization. First, we discuss morphological markers (tumor presence/absence, metastases, subtypes, grades) in which AI-identified regions of interest (ROIs) within whole slide images (WSIs) verifiably overlap with pathologist-identified ROIs. Second, we discuss molecular markers (gene expression, molecular subtyping) that are not verified via H&E but rather based on overlap with positive regions on adjacent tissue. Third, we discuss genetic markers (mutations, mutational burden, microsatellite instability, chromosomal instability) that current technologies cannot verify if AI methods spatially resolve specific genetic alterations. Fourth, we discuss the direct prediction of survival to which AI-identified histopathological features quantitatively correlate but are nonetheless not mechanistically verifiable. Finally, we discuss in detail several opportunities and challenges for these one-label-per-slide methods within oncology. Opportunities include reducing the cost of research and clinical care, reducing the workload of clinicians, personalized medicine, and unlocking the full potential of histopathology through new imaging-based biomarkers. Current challenges include explainability and interpretability, validation via adjacent tissue sections, reproducibility, data availability, computational needs, data requirements, domain adaptability, external validation, dataset imbalances, and finally commercialization and clinical potential. Ultimately, the relative ease and minimum upfront cost with which relevant data can be collected in addition to the plethora of available AI methods for outcome-driven analysis will surmount these current limitations and achieve the innumerable opportunities associated with AI-driven histopathology for the benefit of oncology.

1. Introduction

Histopathology plays a central role in the characterization of biological tissues and is thus a cornerstone of the clinical oncology workflow. Increasingly, whole-slide imaging (WSI) of tissues, along with fast networks data transfer and inexpensive storage, have made it possible to curate large databases of digitized tissue sections [1]. Furthermore, rapid advances in artificial intelligence (AI) methods have enabled scientists to develop automated histopathological analysis methods on WSI, ranging from primitives such as nuclei detection [2] and mitosis detection [3] to more advanced applications such as tumor grading [4]. By and large, these developments provide substantial evidence and great optimism for the future of deep learning as an essential tool for histopathological analyses of WSIs in clinical and biomedical fields.

Despite achieving remarkable success in solving various challenges in histopathological analyses [1,5], AI methods' full potential has been hindered by the perpetual need for painstakingly annotated nuclei, cells, and tissue structures [6–10]. Several recent review articles highlight this specific challenge and limitation for AI methods in histopathology, whether it be for purported clinical utility [11], algorithm improvement [12], or clinical validation [13]. Another recent review article found that across a broad range of histopathological applications, 106/127 studies required some degree of pixel-level annotations for the development of AI methods. Annotations are a limiting factor for the development of AI methods in histopathology.

The driving force behind this phenomenon lies in the nature of the development of cutting-edge AI methods in general computer vision applications. Such methods rely on the abundance of large, annotated

* Correspondence to: 486 N. Patterson Avenue, Winston-Salem, NC 27101, USA.
E-mail address: mniazi@wakehealth.edu (M.K.K. Niazi).

https://doi.org/10.1016/j.semcancer.2023.09.006
Received 24 October 2022; Received in revised form 6 September 2023; Accepted 25 September 2023
Available online 11 October 2023
1044-579X/

datasets. For example, recent viral generative AI models for images such as Dall-E 2 and Midjourney rely on datasets with over one hundred million images. Unfortunately, the advantage of large, annotated datasets does not extend seamlessly to computational pathology, as generating annotations on WSIs demands the expertise of skilled pathologists. By comparison, annotating general-purpose images can be accomplished via unskilled crowdsourcing. Even given pathologists' expertise, their clinical duties often take precedence over research, making it difficult to fully leverage their knowledge for creating annotations. Moreover, this process is laborious and subject to significant variability between different pathologists [1,14–16]. This is compounded by the fact that the number of pathologists is staggeringly low, around 14 per million worldwide, 65 per million in the USA, and fewer than three per million in Africa [17]. Adding to the complexity, the nature of medical image datasets is different from that of general-purpose datasets. WSIs, specifically, are incredibly rich in detail and can have gigapixel resolutions, presenting unique challenges that traditional machine learning and deep learning methodologies struggle to handle [1]. As a result, conventional AI methods often prove impractical in the context of histopathological analyses. Annotations are a barrier to the development of AI methods in computational histopathology.

However, the landscape is evolving, and there is a parad9igm shift towards annotation-less methods in recent high-profile publications [18–20]. Rather than relying on annotations for specific structures like nuclei [21], cells [14], or tissues [22], these methods only require *one label* for an entire WSI, a label that describes a WSI overall, such as malignant/benign. The one-label-per-slide paradigm is known as weak labeling. It simply assumes that only a portion of a WSI corresponds to the weak label. For instance, a WSI might be weakly labeled as cancerous, but only a portion of it contains cancer cells.

This weak labeling approach offers several significant advantages. First, weak labels are more accessible, cost-effective, and considerably less labor-intensive compared to generating strong labels through annotations. Additionally, they minimize the variability between different pathologists since tissue-level labels are no longer necessary. Furthermore, weak labeling opens doors to modeling patient-level data, even when annotations are not feasible. Various crucial signals, such as node status, receptor statuses, tumor mutational burden, chromosomal instability, microsatellite instability, and survival, can be derived directly from the tissue of interest [12]. This is because these signals are derived directly from the tissue of interest, whether from another histological stain or molecular assays.

This paradigm shift toward annotation-less methodologies is a game-changer for the medical and research community, revolutionizing the way they approach histopathological analyses. By harnessing the power of weak labels to overcome the limitations of traditional approaches, the field of computational pathology is poised for exciting advancements and transformative discoveries that can have a profound impact on cancer diagnosis and treatment.

AI methods make use of weak labels via an AI paradigm known as multiple instance learning (MIL) [23]. Weak labels are assigned to collections (called bags) rather than individual examples (called instances), as in conventional AI methods. Classification by MIL methods is then performed at the bag level and not the single instance level. The underlying assumption is that each bag with a specific label shares instances with bags of every other label but also possesses instances unique to itself. For our purposes, the "bags" are the WSIs with weak labels. The "instances" are the unannotated, small image crops sampled from WSIs, although alternative adaptations exist [24,25]. These concepts are depicted in an example in Fig. 1.

Weakly supervised MIL (WS-MIL) presents unprecedented opportunities to enhance oncology through methods that require only patient-level clinical data (i.e., weak labels – one label per WSI) and respective H&E WSIs. To support this area of research, we present a summary of various applications of WS-MIL methods to weakly labeled H&E WSIs with no annotations. We divide these methods first based on their degree of verifiability and then by commonly recurring application areas. Following that, we discuss several opportunities for WS-MIL methods, including lowering the cost of research and clinical care, reducing clinician workload, personalized medicine, and unlocking their potential in the discovery of novel imaging-based biomarkers. Finally, we discuss challenges for WS-MIL methods, including explainability, interpretability, technical validation, reproducibility, data availability, computational needs, data requirements, domain adaptability, external validation, dataset imbalances, and commercialization.

We hold the belief that the adoption of WS-MIL methods, in general, will play a transformational role in bringing computational pathology and the clinical needs of oncology. Through the utilization of these

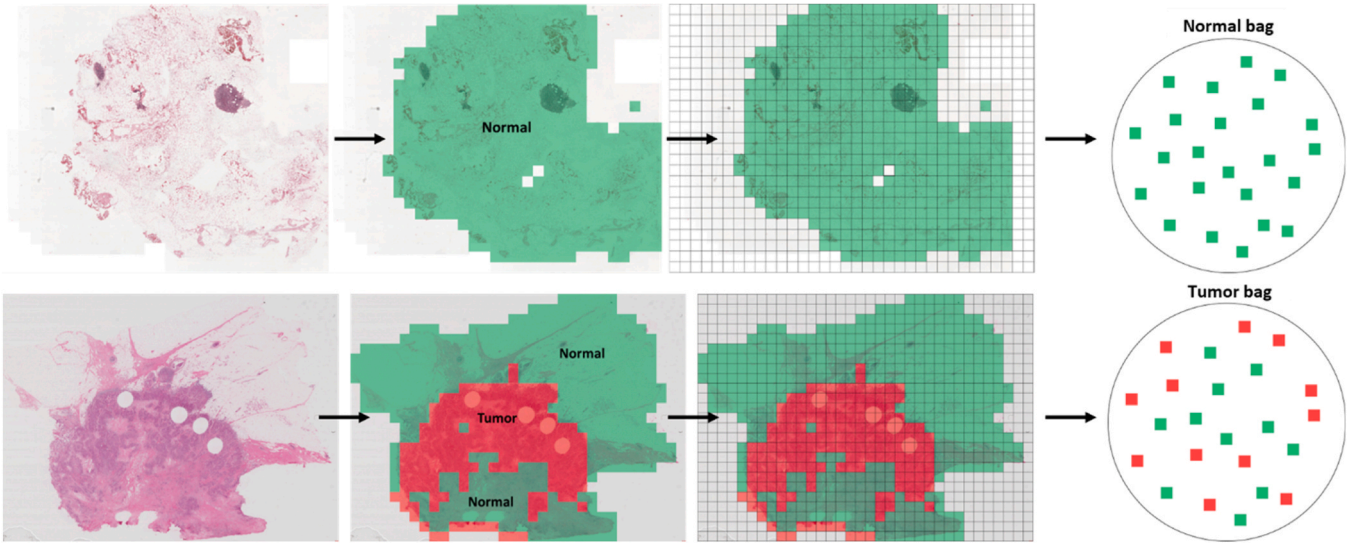

Fig. 1. Example of MIL and weak labeling. Normal WSIs (top row) contain only normal instances (green image crops), and thus we assume that each instance derived from that WSI is implicitly normal. The resulting bag consists of normal instances and is weakly labeled as "normal" for MIL. Tumor WSIs (bottom row) contain both tumor (red image crops) and normal instances, and thus we assume that some instances are implicitly tumor and some are implicitly normal. The resulting bag consists of tumor and normal instances and is weakly labeled as "tumor" for MIL. The green and red annotations are merely for illustration purposes. MIL frameworks do not have access to them.

advanced methodologies mentioned, pathology data, its analysis, and interpretation will become readily accessible, facilitating a seamless integration of expertise between pathologists and oncologists. With WS-MIL, pathology data will be efficiently processed and made available for swift analysis and interpretation. This accessibility will empower oncologists to leverage the rich insights provided by WS-MIL driven analyses in their decision-making. Pathologists' expertise in identifying and characterizing cancerous tissues will be bolstered by WS-MIL's ability to recognize subtle patterns and anomalies, thus enriching the diagnostic process. Oncologists, armed with comprehensive pathology information and AI-generated insights, will be better equipped to make accurate treatment decisions, tailored to the specific needs of each patient. The integration of WS-MIL into pathology and oncology practices holds the promise of elevating patient care. By bridging the gap between these fields, we envision a future where interdisciplinary collaboration becomes the norm, ultimately leading to enhanced diagnostic accuracy, improved treatment outcomes, and, most importantly, better lives for patients battling cancer.

2. Related work

Several reviews have addressed AI methods in cancer histopathology from various perspectives. One early example is a review by Bera et al. [26]. They organize research articles by broad application areas – namely prognostic and diagnostic applications – and based on the AI methods themselves – namely how a modern textbook might categorize AI methods in computer vision. They also focus on the distinction between hand-crafted and automatically learned features. The opportunities and challenges presented by Bera et al. do overlap with some of the same points raised by the current review. For opportunities, Bera et al. mentions enhancing treatment decisions, reducing genomic tests, and biomarker discovery. For challenges, Bera et al. mention interpretability, validation, and clinical potential. These are discussed in the current review but mainly from the perspective of WS-MIL. Likewise, we offer many more discussion points as well as how WS-MIL uniquely relates to them.

A review by Echle et al. [12] focuses on different AI application areas in histopathology, all of which overlap with the current review. These include tumor detection, subtyping, and grading (which they refer to as "basic" applications") as well as mutation prediction, treatment response prediction, and survival prediction (which they refer to as "advanced applications"). They also divide research articles by their clinical readiness – internal validation, external validation, and finally FDA approval – and primarily focus on clinical implementation. In comparison, the current review presents more application areas (gene expression, molecular subtyping) and more granular divisions of tumor detection and mutation prediction. Moreover, the current review focuses on WS-MIL and the opportunities and challenges it uniquely presents, though some of the research articles Echle et al. cite happen to use such methods. Finally, the current review stratifies WS-MIL methods by the ability of researchers and pathologists to visually validate and verify them.

The more recent review by Kleppe et al. [27] focuses primarily on study design and external validation. They discuss how current applications of AI methods to medicine tend to lack robust study design and those that do rarely validate on external cohorts. The current review does discuss these concepts as pervasive issues as.

they related to WS-MIL, namely the subsections on reproducibility and homogeneity of data, but for readersinterested in a more in-depth dive into these issues, we refer them to the review by Kleppe et al. [27]. Similarly, a review by Cifci et at al. [105] discusses AI methods to predict genetic alterations, discussed in-depth in the current review. However, the current review focuses primarily on WSI-MIL and offers a unique perspective on the verification of WS-MIL methods to predict genetic alterations. For readers interested in a deeper dive into applications to genetic alterations, we refer them to Cifci et al. [105].

Finally, a recent review by Shmatko et al. [106] discusses some of the same concepts as in the current review. For example, they cite examples of various studies that have developed models for tumor detection, subtyping, grading, survival, treatment response, risk scores, molecular markers, gene expression, and genetic alterations (MSI specifically). Some of the studies they cite when discussing these concepts also overlap with the studies cited in the current review. However, when discussing these application areas, they also cite studies that do not apply WS nor MIL. No distinction is drawn. Still, Shmatko et al. do have a subsection of their "Methodological innovations" section that discusses WS and MIL. This is a small portion of their review, and they reference only five articles related to MIL while only superficially summarizing them. The current review has 91 WS-MIL articles and goes into more detail for many studies. Beyond that, the current review structures these application areas in a unique manner. Namely, how WS-MIL methods are or are not verifiable in the sense that pathologists can point to exact regions that correspond to an overall slide-level label (i.e., morphological or molecular markers) but sometimes cannot (genetic markers or survival). Finally, the challenges and opportunities presented in the current review article are unique to WS-MIL, whereas the review article by Shmatko et al. is general to computational pathology within oncology. We want to bring focus to WS-MIL, as the vast majority of studies still rely on annotations when they are simply not necessary. WS-MIL has general applicability and is relatively easy to apply to many problems within computational pathology.

3. Methods

Current AI methods that model patient-level clinical data from H&E WSIs can be characterized in terms of their interpretability and explainability. Interpretability refers to the ability of a model to localize ROIs in a WSI that drive its overall prediction. For example, a model may predict that a given WSI has a tumor. If the same model localizes an ROI that drives that overall prediction, it can be said to be interpretable. Explainability refers to whether those ROIs can be explained by humans. For example, consider the previous model. If a pathologist can meaningfully interpret the ROI as correlated to tumor presence, then that model is said to be explainable. Both characteristics ultimately impact clinical viability. These characteristics form a basis for *visual verification and validation*. We demonstrate that WS-MIL applications to H&E WSIs exist on a continuum in terms of the degree to which they can be visually validated or verified (see Fig. 2). It is worth noting that WS-MIL is simply a general approach from which many specific methods are derived. For the sake of brevity, specific methods are not discussed and instead referred to as simply "WS-MIL." Additionally, the terms "method" and "model" are used interchangeably but refer to the same concept – a deep learning algorithm. For interested readers, all referenced WS-MIL methods are organized by application in Table 1.

3.1. Morphological markers: verifiable WS-MIL with pathologist ROIs

The main characteristics of these methods are that 1) they attempt to predict something about tissue morphology and 2) they are relatively easy to verify, as a pathologist can point to a region of interest driving the overall weak label while WS-MIL methods can highlight which regions of the WSI drive its overall decision. Thus, if the pathologist and WS-MIL ROIs overlap, one can verify that the model attends to the correct regions of the WSI. Fig. 3 depicts this verification process. Common application areas within this category include cancer presence or absence, subtyping, grading, and site of origin prediction. Automation of such tasks is relevant, as they can be laborious and time-consuming for pathologists, may allow for triaging or selection of slides that may or may not contain tumors (thus reducing pathologist workload), and may serve as a second opinion (thus reducing errors). Any derived benefits for pathologists would ultimately benefit oncology.

For cancer presence or absence, the task is to predict if a WSI

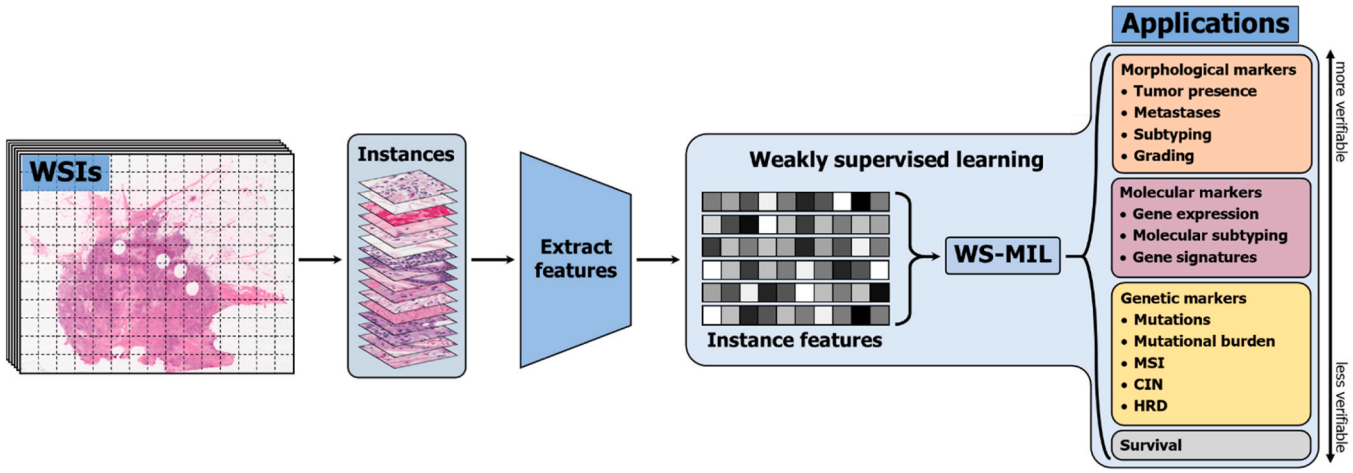

Fig. 2. Variably verifiable categories of WS-MIL model targets – morphological, molecular, and genetic markers – and specific application areas. WSIs are first split into instances (small image crops). Then, features are extracted from each instance individually. Such features are either automatically learned by the WS-MIL model during training or alternatively pre-determined using a feature extraction model trained on a different dataset. This results in a set of features for each instance (instance features). Such features may or may not be interpretable by humans but nonetheless can be utilized by a WS-MIL model. Finally, WS-MIL processes all instance features to yield a WSI-level output.

contains a tumor. This also includes the prediction of metastasis presence or absence, as metastases are also tumors. The preliminary work to tackle this task using WS-MIL was done by Courtiol et al. in 2018 [48]. They applied their WS-MIL method to a commonly utilized dataset, Camelyon16 [16], which consists of sentinel lymph nodes that are either present or absent in breast cancer metastases. They demonstrated that WS-MIL could achieve a high area-under-the-curve (AUC) and that their model attends specifically to pathologist-annotated metastases (rather than arbitrary tissue regions). This was notable at the time, as previous methods had relied on tediously annotated tissue ROIs [16]. Several studies have since followed suit, increasingly applying WS-MIL to Camelyon16 and observing the qualitative and quantitative overlap between model-attended instances and pathologists' annotations [4,18,42–47,50–61,107]. Other studies have examined lymph node metastasis on the smaller MSKCC breast cancer dataset [49] and in-house datasets of many thousands of WSIs [75] or a few hundred for colorectal cancer [77]. The main limitation of these methods (and perceptively deduced by Courtiol et al.) is that metastases are highly localized and restricted to a small area within the WSI. Thus, it is difficult to develop a method that can accurately identify tumor *instances* (as in Fig. 1) when they may only constitute $< 0.001\%$ of the total tissue area. In other words, the signal of metastasis is weak and thus hard to detect. It is therefore interesting when one study sought to predict metastases from primary tumors [78] from multiple sites-of-origin and another study solely for the bladder [30], completely circumventing this problem of small tumors.

This is contrasted with more generalized cancer presence or absence tasks in which the diseased tissue is larger and more diffuse. Campanella et al. are often cited as the first study to apply WS-MIL to this task [4]. In addition to Camelyon16, they applied their method to three large sets of WSIs including an in-house prostate cancer dataset ($n = 12{,}132$), an external prostate cancer dataset ($n = 12{,}727$), and a skin cancer basal cell carcinoma dataset ($n = 9962$). Their results showed that WS-MIL on large datasets outperforms strong supervision on small datasets. This was significant, as previous studies relied on pathologists' annotations of tumor regions in small WSI datasets (with the obvious exception of earlier work by Courtiol et al.). Several studies have similarly applied WS-MIL for tumor presence or absence classification in WSIs and likewise have observed the qualitative and quantitative overlap between model-attended instances (by the MIL model) and pathologists' annotations [42,76,82,93,96,103,104,107,108] in breast, lung, stomach, colorectal, thyroid, liver, cervical, and prostate cancers. One study in particular by Fu et al. trained a model across 42 tissue types, achieving AUCs above 0.95) for all tissue types [29].

Like the cancer presence or absence task, subtyping tends to involve cancers in which the diseased tissue is larger and more diffuse. Courtiol et al. [48] were the first to apply WS-MIL to subtype cancer using WSIs – specifically, subtyping non-small-cell lung cancer (NSCLC) into lung adenocarcinoma (LUAD) and squamous cell carcinoma (LUSC) using WSIs available from The Cancer Genome Atlas (TCGA) [109] – and drew similar conclusions and implications from their work with Camelyon16. Not surprisingly, the standard set by Courtiol et al. has endured through several subsequent studies subtyping NSCLC into LUAD and LUSC using TCGA [18,43,46,47,50,52,54,57,59,62–64,96,98,99,107] as well as augmenting the dataset with multi-institutional WSIs [97]. Wang et al. added a subtype for small-cell lung cancer (SCLC) [100] using an in-house dataset, while Zheng et al. added a category for cancer-free tissue [84]. An earlier study by Hou et al. also looked at NSCLC but using an earlier version of TCGA [36]. Beyond lung cancer, other subtyping tasks have also been studied using WS-MIL, most popularly renal-cell carcinoma subtyping into papillary (PRCC), chromophobe (CRCC), and clear cell (CCRCC) [18,43,50,52,55,59,62–64]. Several studies have also applied WS-MIL to subtype breast cancer into invasive ductal carcinoma (IDC) and invasive lobular carcinoma (ILC) [58,62–64]. Many additional one-off studies have applied WS-MIL to various cancer subtyping problems. Hashimoto et al. subtyped lymphomas into diffuse large B-cell (DLBCL) and non-DLBCL [41]. Specifically, they lumped germinal center B-cell (GCB) and non-GCB into DLBCL and angioimmunoblastic T-cell (AITL), Hodgkin's mixed cellularity (HLMC), and Hodgkin's nodular (HLNS) into non-DLBCL. Lu et al. subtyped gliomas into glioblastoma (GBM) oligodendroglioma (O), and astrocytoma (A) while ignoring oligoastrocytoma (OA) [35] as did Hou et al. [36]. Li et al. also examined the subtyping of various primary tumors, including five glioma subtypes – diffusive astrocytoma (DA), anaplastic astrocytoma (AA), O, and anaplastic oligodendroglioma (AO) – and six meningioma subtypes (fibrous, meningothelial, transitional, angiomatous, atypical, and anaplastic) [37]. The aforementioned study by Zheng et al. additionally examined the computational efficiency for various WS-MIL approaches through the subtyping of 6 types of gastric pathology – including low-grade intraepithelial neoplasia (LGIN), high-grade intraepithelial neoplasia (HGIN), adenocarcinoma (A), mucinous adenocarcinoma (MA), Signet-ring cell carcinoma (SRCC), and normal tissue – and subtyping 5 types of endometrial pathology – including well/moderately/low-differentiated endometrioid adenocarcinoma, (WDEA/MDEA/LDEA), serous endometrial, intraepithelial carcinoma

Table 1
Studies cited in this review organized by organ/system in rows and application areas (columns).

	Morphological markers				Molecular markers			Genetic markers					Clinical outcomes	
	Tumor presence	Metastases	Subtyping	Grading	Gene expression	Molecular subtyping	Gene signatures	Mutations	Mutational burden	MSI	CIN	HRD	Survival	Treatment response
Adenoid					[28]									
Adrenal	[29]	[19]			[29]			[29]						
Bile duct					[28]									
Bladder	[29]	[19,30]			[28,29]			[29]					[29,31–34]	
Brain		[19]	[35–37]		[28,29]			[29,37,38]					[29,32,33, 39,40]	
Blood		[19]	[41]		[28]	[41]								
Bone					[29]									
Breast	[29,42]	[4,18, 42–61]	[58,62–64]		[28,29,65, 66]	[62,67–70]	[68]	[29,38,65, 68]			[71]	[68]	[29,31–34]	[72]
Cervical	[29]	[19]			[28,29]			[29,68]					[29]	
Colorectal	[29,73,74]	[19,75–78]		[73]	[28,29]	[68]	[68]	[29,38,68, 79,80]	[68,80]	[78, 80–82]	[80]	[81]	[31,34,40, 83]	
Endometrial		[19]	[84]					[38,85]						
Esophageal	[29]	[19]	[59]		[28,29]			[29]					[29]	
Gastric	[29,42]			[84]	[29]	[68]	[68]	[68]	[68]	[68]		[68]	[29,31,34]	
Germ cell		[19]												
Head/neck	[29]	[19]			[28,29]	[86]		[29,68]					[31,34]	
Kidney	[29]	[19]	[18,43,50]	[44,52,55,59, 62–64,87–89]	[28,29,90, 91]			[29,68]					[29,31,34, 92]	
Liver	[93]	[19]			[28,29]			[29,68,94]					[34,92,95]	
Lung	[29,42,96, 97]	[19]	[18,36,43,46–48,50,52, 54,57,59,61–64,84, 98–100]		[28,29]	[68]	[68]	[29,38,68, 98]				[68]	[31–34,39, 83,92]	
Mesothelium	[29]				[28,29]			[29]					[101]	
Neuroendo					[28]									
Ovarian	[29]	[19]			[28,29]			[29]					[29,31]	
Pancreas		[19]			[28]			[68]					[34]	
Prostate	[4,29]	[19]		[15,102,103]	[28,29]			[29,68]						
Skin	[4,29]	[19]			[28,29]			[29,68]					[29,34]	
Testicular	[29]				[28,29]			[29]						
Thyroid	[29,104]	[19]			[28,29]			[29]						
Uterine	[29]				[28,29]			[29]					[29,32–34]	

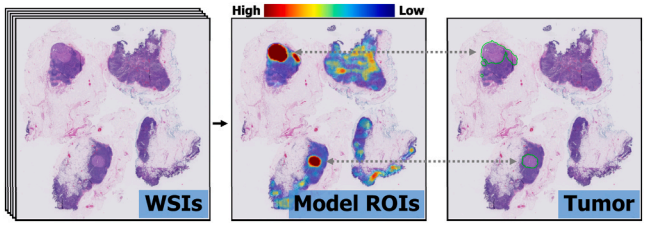

Fig. 3. Morphological markers verification – pathologists can point to an ROI driving the weak label while MIL can highlight which regions of the image drive its overall prediction. Tumor presence is given as an example.

(SEIC), and cancer-free tissue [84]. In a similar fashion to the original Courtiol et al. study, Zhu et al. applied WS-MIL to classify esophageal cancer WSIs from TCGA into two subtypes – adenocarcinoma and squamous cell carcinoma [59].

Finally, WS-MIL has been applied to grading – most popularly, Gleason grading [110]. Bulten et al. [15] removed the need for manual annotations by utilizing two pre-trained models to 1) delineate a rough tumor outline and 2) remove epithelial tissue from WSIs. Then, all tissue regions were labeled with the pathologist's reported Gleason pattern. During the training of their model, only "pure" biopsies were included (i.e., 3 +3, 4 +4, 5 +5). They showed that despite this weak labeling strategy, their method was able to accurately score Gleason grades *outside* the original domain (i.e., 3 +4, 4 +3, etc.) and had a high consensus with experts. Subsequent studies have also applied WS-MIL for Gleason grading [88,102,103,111], grading dysplasia of various cancers (i.e. normal, low-grade, high-grade, cancer) [73,74], grading of colorectal cancer (i.e. low-grade and high-grade) [82], and Fuhrman grading of CCRCC [89].

3.2. Molecular markers: WS-MIL verifiable with adjacent tissue

The main characteristics of these methods are that 1) they attempt to predict something about molecular markers of cancer and 2) they are relatively difficult to verify, as a pathologist cannot point to a region of interest driving the overall weak label. While WS-MIL can still highlight which regions of the image drive its overall decision, physical methods are required to verify whether these ROIs (such as adjacent IHC, immunofluorescence, or digital spatial profiling) correspond to the overall weak label. Thus, if positive regions of adjacent tissue and WS-MIL ROIs overlap, one can verify that the model attends to the correct regions of the WSI. Fig. 4 depicts this verification process. Common application areas within this category include gene expression and molecular subtyping. Such tasks are of interest, as H&E is relatively ubiquitous, inexpensive, and takes less time than molecular assays.

For gene expression prediction, the primary task is to *regress* gene expression values directly from H&E WSIs. These values are obtained either via RNA-seq or via microarray. The seminal study that tackled this task using WS-MIL was Schmauch et al. in 2020 [28]. Their model

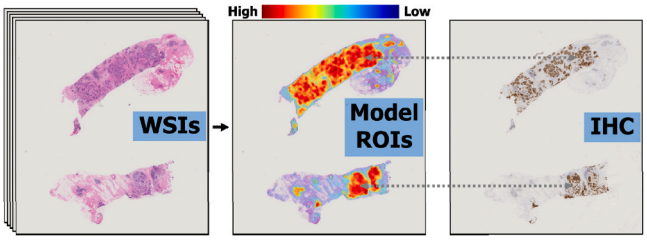

Fig. 4. Molecular markers verification – pathologists cannot point to an ROI driving the overall weak label, but physical methods (such as adjacent IHC) can help verify whether these ROIs correspond to ROIs highlighted by WS-MIL. HER2 scoring is given as an example. Such H&E features identified by WS-MIL tend to be beyond human comprehension.

predicted RNA-seq-derived expression from H&E with a significant correlation for 28% of genes and 33% of protein-coding genes. Gene enrichment analysis also revealed that their model could accurately predict dysregulated pathways. Moreover, they were able to apply their model to resolve genes from H&E spatially. More significantly, they verified spatially-resolved CD3 and CD20 on an external dataset of adjacent H&E and IHC WSIs *and* spatially-resolved epithelium-associated genes on an external dataset of H&E WSIs annotated for epithelium by pathologists. A study of a similar scale by Fu et al. predicted gene expression for a wide range of cancer types across 17256 genes and achieved a correlation above 0.25 for 42% of genes [29], though histological subtypes and grades achieved similar distributions of correlation values across all genes. Weitz et al. [66] similarly showed a significant correlation between RNA-seq-derived expression and model-predicted gene expression for approximately ~89% of genes on internal and external testing sets of H&E WSIs. Spatially-resolved predictions (as determined from GeoMX data) were significant for 24% of genes. Freyre et al. spatially resolved Kim-1 protein and mRNA models using H&E and qualitatively confirmed findings using adjacent in-situ hybridization [90]. Finally, Alsaafin et al. were able to accurately predict RNA-seq values from H&E WSIs in such a manner that the WSI representations could be used for downstream tasks such as subtyping and retrieval [91].

Unlike regressing gene expression directly, in molecular subtyping, the aim is to *classify* WSIs based on the presence, absence, or degree of a particular molecular marker. Such molecular markers may include over/under-expressed receptors, abnormal methylation patterns, or specific cell populations. For example, automated HER2 scoring based on IHC is one of the more popular applications of deep learning models to pathology [112]. Here, the task is to assign a score of 0, 1 +, 2 +, or 3 + to pre-selected WSI ROIs (i.e., not weakly supervised). Subsequent studies have accomplished automated HER2 scoring in breast cancer with WS-MIL from H&E, one without qualitative or quantitative validation on adjacent IHC [69], and one with validation [70]. Likewise, Naik et al. applied their weakly-supervised ReceptorNet to predict HER2 and PR positivity in the H&E WSIs [67]. Additionally, they demonstrated that their model was statistically agnostic to the data source, institution, menopausal status, or race (a common shortcoming of WS-MIL methods in histopathology). Finally, they demonstrated how their model discovers histomorphological patterns important for HER/PR receptor status by asking pathologists to interpret model-selected ROIs. Apart from breast cancer, other studies have examined predicting MYC rearrangements in DLBCL [56] and scoring LMP1 in nasopharyngeal carcinoma [86], from H&E using WS-MIL, although neither performed any quantitative or qualitative validation of their results. Finally, one study examined an application of WS-MIL to predict MGMT methylation status in gliomas and found qualitative overlap between attended ROIs and adjacent IHC [37].

Distinct from the prediction of receptor status, the earliest application of WS-MIL to molecular subtyping was classifying lymphomas into DLBCL and non-DLBCL [41], as described earlier. Though the target of the study was not explicitly *molecular* subtyping, DLBCL and non-DLBCL are differentiated by the presence of B-cells (which express CD20). Thus, the implicit target of their study was the molecular subtyping of lymphoma. Strikingly, they showed that though their model used only H&E WSIs, model-attended regions corresponded to CD20 positivity (as confirmed by adjacent IHC). This intuitively confirmed that their model was able to identify B-cells (differentiated by CD20 expression) while only knowing weak DLBCL and non-DLBCL subtypes.

A large-scale, externally-validated study by Kather et al. [68] performed classification on a wide array of cancers and molecular subtypes including lung adenocarcinoma (TGF-β, CD8, LUAD-3, LUAD-4, LUAD-5, LUAD-6), colorectal (Wnt/β-catenin, CMS1, CMS2, CMS3, CMS4), breast (HER2, ER, PR, TGF-β, PAM50, LumA, LumB), and gastric (TGF-β, CIMP-L). Their study showed that all of these molecular subtypes could be accurately predicted from H&E WSIs alone. Furthermore,

they validated model-attended regions with pathologist interpretation on H&E. Finally, their model was unified across all sites of origin, therefore generalizing to all cancers studied.

3.3. Genetic markers: WS-MIL that goes beyond human comprehension

The hallmarks of these methods are that 1) they predict some genetic markers of cancer and 2) they are not possible to verify, as a pathologist cannot point to a region of interest driving the overall weak label. Furthermore, unlike molecular markers, there are no methods to validate if regions that WS-MIL methods highlight truly correspond to the weak label, although several studies visualize these model-attended ROIs to no avail. Fig. 5 depicts this process. Common application areas within this category include mutation prediction, mutational burden, microsatellite instability (MSI), chromosomal instability (CIN), and homologous recombination deficiency (HRD). Like with molecular markers, automated detection of such markers via H&E is popular, as the required physical assays are expensive, tedious, and time-consuming, whereas H&E is inexpensive, part of standard clinical workflow, and could potentially be augmented with WS-MIL with a simple click of a button. It is worth noting that several of these studies reported histomorphological features of model-attended ROIs, which in some cases were known to correlate with the presence of the genomic target of interest. However, it is currently not possible to spatially resolve cells with specific genetic alterations; thus it cannot be confirmed if model-attended ROIs truly correspond to tumor cells with said alterations.

For mutational prediction, the task is to predict whether a particular gene is mutated based on an H&E WSI. The earliest work to study mutation prediction via WS-MIL applied to H&E was by Coudray et al. in 2018 [98]. Not only did their model subtype lung cancer into LUAD and LUSC, but it also accurately predicted the presence of ten lung cancer-associated genes with AUCs ranging from 0.64 to 0.85, all with statistical significance. Liao et al. [94] generated similar results for hepatocellular carcinoma with AUCs ranging from 0.52 to 0.90 for ten cancer-associated genes, most statistically significant. Unlike the earlier study, they were able to validate their model on TMAs though only trained on WSIs from TCGA. Perhaps the largest study was carried out by Kather et al. (mentioned previously) [68] and involved examining multiple cancers (LUAD, colorectal, breast, gastric, melanoma, prostate, pancreatic, LUSC, hepatocellular, PRCC, CRCC, CCRCC, head and neck, and cervical) and 95 cancer-associated genes (across all sites of origin). Critically, they distinguished between all mutations and oncogenic drivers, as not all genetic variants of cancer are causative of malignant processes. Their pan-cancer model resulted in AUCs of 0.3–0.8. When limited to oncogenic drivers, the upper range of AUCs increased to 0.95 probably due to less noise attributable to non-oncogenic drivers. In an almost identical fashion, another study predicted mutations associated with breast cancer [65] with AUCs reported for 6/18 genes ranging from 0.68 to 0.85. Fu et al. were able to achieve AUCs ranging from 0.6 to 0.9 for oncogenic drivers across a wide range of cancers [29]. And finally, another large-scale study examined a vast array of genomic markers via WS-MIL in colorectal cancer and achieved AUCs of 0.79, 0.73, and 0.60 for BRAF, TP53, and KRAS [80]. Rather myopically, the authors indicate their model outperforms Kather et al. [68] even though the model by Kather et al. spanned many sites of origin while the authors' focused on just one and therefore does not merit a direct comparison. Furthermore, they were outdone by subsequent studies. One that achieved AUCs of 0.82 and 0.61 for BRAF and KRAS [82]. The other achieved AUCs of 0.82 and 0.67 for BRAF and KRAS [79], in addition to predicting NRAS and PIK3CA in colorectal cancer. The same group performed a subsequent study examining the generalizability of their model to pan-cancer external cohorts [38]. One study greatly improved on the prediction of LUAD mutations, albeit without direct comparison to Kather et al. [51]. Another study examining solely gliomas (which Kather et al. did not) achieved AUCs of 0.8975 and 0.8175 for predicting IDH1 and TP53 mutations, respectively [37]. Finally, Fremond et al. leveraged several clinical trial datasets using WS-MIL to not only accurately predict mutations in endometrial cancer but also qualitatively correlate model-attended ROIs to morphological and cellular features [85]. These included a high density of lymphocytes for POLE mutation, strong nuclear atypia for P53ABN, and inflammatory morphology and MMRd (mismatch repair deficiency).

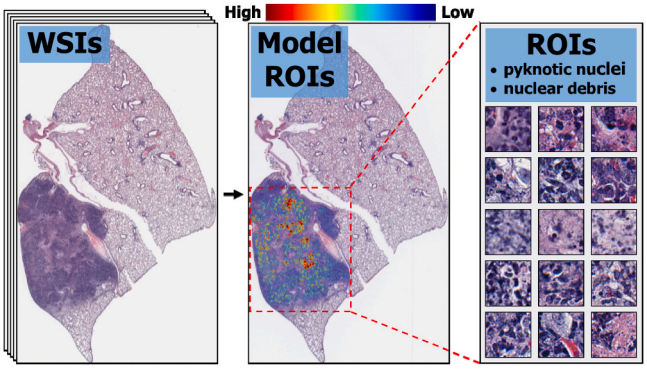

Fig. 5. Genetic markers and survival verification – pathologists cannot point to an ROI driving the overall weak label, and no physical methods can validate it as in molecular markers. However, pathologists may characterize said model ROIs. Survival prediction is given as an example. Such H&E features identified by WS-MIL tend to be beyond human comprehension.

For mutational burden, the task is to predict a high degree of somatic hypermutation of tumor cell DNA from H&E images alone. Only two studies have addressed this task using WS-MIL (and have already been mentioned previously). First, Kather et al. [68] showed that for colorectal and gastric cancers, the mutational burden could be predicted with AUCs of 0.71–0.75, respectively, with significance. By comparison, Bilal et al. [80] achieved an AUC of 0.90 on the same cohort, albeit under a simpler experimental setup (see above paragraph). Interestingly, the author's [80] method provides some automated insight (as opposed to pathological interpretation) into the contributions of various histological features and cellular compositions that contribute to mutational burden according to their WS-MIL model – inflammation and neoplastic epithelium type 1.

MSI refers to the degree to which repeated elements in DNA change in length. Kather et al. were the first to predict MSI status using WS-MIL methods, albeit only a single data point in a tour-de-force of various molecular and genetic markers for multiple types of cancer. Specifically, they achieved an AUC of 0.61 for a gastric cancer dataset (though their model was trained on several cancers and outcomes) [68]. Bilal et al. focused their study on only colorectal cancer and achieved an AUC of 0.90 [80]. Schrammen et al. [82] developed a method called SLAM that significantly outperformed all previous state-of-the-art WS-MIL methods for H&E WSI analysis achieving an AUC of 0.900 on an external dataset of colorectal cancer cases. Although model-attended ROIs could not be verified, they did show that non-ROIs corresponded to normal tissue, which is expected. This was yet again outperformed by Niehues et al., who achieved an AUC of 0.92 on a large, externally validated dataset [79]. Finally, the WS-MIL model by Brockmeller et al. achieved an AUC of 0.793 in pT2 colorectal cancer H&E WSIs [78].

CIN refers to the degree to which chromosomes or sections of them are duplicated or deleted. As in MSI status, Kather et al. [68] first addressed CIN using WS-MIL. Their model achieved an AUC of 0.73 for colorectal cancer. Again, Bilal et al. [80] were able to improve on this for the same cohort using their method and achieved an AUC of 0.85. Most recently, Xu et al. were able to achieve an AUC of 0.822 for a cohort of breast cancer cases [71]. Interestingly, they showed that there was no statistically significant difference in model performance across breast cancer molecular subtypes (ER, PR, HER2). This was important, as CIN was more prevalent in ER-, PR-, and triple-negative subtypes for their dataset.

HRD refers to the loss of the ability of cells to repair double-stranded DNA breaks through homologous recombination. As with other genetic markers, Kather et al. [68] applied WS-MIL to predict HRD in LUAD, breast cancer, and gastric cancer. Given that HRD is a continuous value, Kather et al. binarized HRD into two categories – HRD-low and HRD-hi – relative to the overall median. They achieved AUCs ranging from 0.66 to 0.76. Schirris et al. [81] and Lazard et al. [62] improved marginally upon these results for breast cancer.

3.4. Survival: WS-MIL that correlates but cannot be verified

For survival, the task is to *regress* the number of days a patient survived (as all data is retrospective). Like with genetic markers, there is no way to validate whether WS-MIL identified ROIs correspond to specific survival times. Likewise, there are histopathological features that correlate with survival but nonetheless cannot be tied directly to specific survival times. Related to survival is *recurrence*, however no application of WS-MIL to H&E WSIs has been reported. Fig. 5 depicts this process.

WSISA was the seminal WS-MIL method to predict survival from H&E WSIs [39]. They tested several state-of-the-art survival models in conjunction with their WS-MIL method and compared them to ROI-based methods. Across two lung datasets (NLST, TCGA-LUSC) and one glioblastoma dataset (TCGA-GBM), their results indicated that their WSI-based method outperformed ROI-based methods. The same group developed another WS-MIL method [113] which reportedly outperformed their previous method. Strangely, they did not report the same performance metrics for WSISA for the same datasets on this follow-up study. However, they qualitatively showed that their model attended to ROIs annotated by pathologists (without being explicitly trained to do so).

Two subsequent studies [95,101] predicting survival in mesothelioma and hepatocellular carcinoma showed that their WS-MIL methods significantly outperformed those multivariate methods based on all available clinical, biological, or pathological variables. For both studies, automatically identified ROIs predictive of low survival were vascular spaces and pleomorphisms whereas for high survival were inflammation and immune infiltration. A study by Google presented results for their survival prediction method across ten TCGA datasets, although failing to compare to any previous methods and leave model-attended ROIs up to reader interpretation [31]. Another study built on and compared their method (DeepAttnMISL) to the work by Zhu et al. [39] and Li et al. [113] and achieved improved survival prediction performance for TCGA-LUSC [83].

This was surpassed by yet another WS-MIL method by Chen et al. which combined WSIs with genomic data for several cancer, including bladder (TCGA-BLCA), breast (TCGA-BRCA), brain (TCGA-GBM), lung (TCGA-LUAD), and uterine (TCGA-UCEC) [33]. Their model was unique in that cellular processes (represented by genomic data) could be visualized directly on H&E, such as tumor suppression, oncogenesis, protein kinases, cellular differentiation, transcription, and cytokines. Chen et al. developed another method to predict survival [32] which outperformed their previous method on TCGA-GBM. Across all cancers, their model attended to necrosis, dense tumor aggregates, and regions of desmoplastic stroma, indicative of tumor invasion and proliferation. Shao et al. [92] proposed yet another WS-MIL method for survival prediction which outperformed several previous methods (though Chen et al. [32,33] were not reported) for kidney (TCGA-KIRC), liver hepatocellular carcinoma (TCGA-LIHC), and TCGA-LUSC cancer datasets. Chen et al. [40] in their third method to predict survival added an additional component to their WSI/genomic model which integrated spatial information of WSIs, which significantly outperformed their original model for TCGA-GBM and TCGA-KIRC. Finally, Chen et al. extended their third work to include 14 cancer types, definitely demonstrating not only that multi-modal approaches to predict outcomes perform more accurately than single modes but that such approaches can be leveraged to discover prognostic features that correlate with these outcomes [34].

4. Opportunities

The paradigm of driving model development through the utilization of H&E slides to model clinical data creates several opportunities within research and clinical care. Thus far, we have described how these methods can be categorized in terms of their degree of *verifiability*. All studies discussed thus far highlight their methods' verifiability, interpretability, and explainability by selecting and analyzing model-attended ROIs and comparing them to some source of ground truth. These features present clear opportunities for WS-MIL in oncology. Their ubiquity is due to their inherence within MIL. Given the focus that verifiability, interpretability, and explainability have been given thus far, the following Section will discuss further opportunities including but not limited to reducing the cost of research and clinical care, reducing the workload of clinicians, personalized medicine, and finally unlocking the full potential of histopathology.

4.1. Reducing the cost and expanding the scope of research

As has been repeatedly stated, WS-MIL only requires one label per slide. Therefore, there is little to no need for pathologists to annotate WSIs for WS-MIL applications. This reduces the cost of research not only in terms of labor but also time, given that annotations are laborious and time-consuming [4,31,32,41,46,49,74,76,90,93,96,99,100]. Furthermore, because of the lack of need for annotations, the number of slides that are integrated into model development is limited not by the amount of labor available but instead by the total number of slides with a weak label. In fact, several studies have shown that WS-MIL on large datasets tends to lead to higher generalization performance (i.e., performance on an external cohort) than fully-supervised on small, curated datasets [4,96]. This is because the diversity of WSIs in terms of staining protocols, scanners, and disease manifestation is better represented by larger datasets. Removing the annotation burden reduces the cost of research.

Even if annotations are needed, WS-MIL offers methods to speed up the process of obtaining annotations [48] through methods such as *attention*. In the former, a model can be trained with some weak labels associated with ROIs. For example, if one were interested in annotating tumor regions, only a weak label for WSI (i.e., containing tumor or not containing tumor) would be needed. Then, after training, model-attended ROIs would tend to correspond to tumor regions, which could then be automatically annotated by the WS-MIL (and subsequently edited by pathologists). Several studies have demonstrated this potential [76,82,93,96,104]. Reducing the annotation burden reduces the cost of research.

Finally, WS-MIL enables the prediction of genetic markers and the quantification and localization of molecular markers that would otherwise be impossible to annotate or at least would be expensive to annotate [67,90]. We have presented how WS-MIL can be applied to predict genetic markers such as mutation prediction, mutational burden, MSI, and CIN. These are not just impossible to annotate but require expensive tests. WS-MIL could offer a low-cost alternative for research applications. Similarly, we've shown that WS-MIL can model and spatially resolve molecular signals such as gene expression and subtypes. Such signals could be annotated on H&E through adjacent molecular stains (IHC/IFC) or spatial transcriptomic methods through image analysis methods. However, such analyses are quite expensive. WS-MIL not only provides a low-cost virtual alternative that only requires H&E, but it also reduces the need for tissue [19].

4.2. Reducing the cost of clinical care

Specialized molecular assays are costly and take time to turnaround in the clinic. The majority of studies described in this review argue that relative to such assays, H&E WSIs stand inexpensive and routine in the oncological workflow. Furthermore, these studies assert the assumption that H&E WSIs contain a rich source of information ripe for WS-MIL

enhanced methods to exploit for the virtualization of such molecular assays. For example, Naik et al. argue and provide results that suggest that the need for IHC may be reduced by virtually re-straining H&E or by approximating the target of interest with H&E both by utilizing WS-MIL driven methods [67]. Lu et al. similarly suggest that molecular tests for determining cancers of unknown primary (or whether a tumor is primary or of metastatic origin) may be supplanted by WS-MIL methods applied to H&E. This is not too farfetched, given that their model was able to perform these tasks across 17 cancer times with high accuracy [19]. Virtualization of such methods would not only reduce cost but would also reduce turnaround time in cases requiring molecular profiling, [80], arduous analysis [28], extra staining of tissues [19,48, 67,80,98], and just in general [100]. Inter and intra-observer variability in the interpretation of IHCs would also be greatly minimized [46,67]. Lastly, resource-constrained settings would also benefit [19] in which fewer clinical and ancillary tests are available [19,96].

In addition to the potential time-cost benefits of virtualizing molecular tests, WS-MIL on H&E WSIs may also reduce the cost of clinical care through patient triaging [18,96]. For example, Campanella et al. [4] demonstrated how their WS-MIL method (which simply detects whether an H&E WSI contains a tumor) could potentially reduce the number of slides reviewed by pathologists by 75%. In their hypothetical diagnostic scenario, their WS-MIL method would triage prostate WSIs according to the probability that they contained tumor. Then, pathologists would review cases more likely to contain tumor, essentially disregarding all benign slides. In other words, their WS-MIL method would screen WSIs for containing primary tumor while retaining 100% sensitivity, at the cost of specificity. This is especially important for applications in prostate cancer, as several WSIs are taken per patient, and only a few contain tumor to be further analyzed. Beyond triaging slides, patients may also be stratified based on AI-predicted risk. Several studies have shown that patient survival can be accurately predicted from H&E. Further, studies have even shown that WS-MIL improves survival prediction over clinical standards [40,83,92]. Such refinement of the prediction of patient prognosis could potentially improve treatment allocation [95] and reduce inter-and intra-observer variability [67] thus reducing errors and improving overall patient outcomes.

4.3. Reducing the workload of clinicians

Successful enhancement of pathology workflows through WS-MIL has a potential to decrease clinician workload through patient stratification as discussed previously. Specifically, Kaplan-Meier curves significantly separate long- and short-term survival based on WS-MIL predicted survival and significantly outperform current clinical models [18,40,92,96]. Moreover, WS-MIL will decrease clinician workload by functioning as a second opinion [19,41,98]. Several studies have shown that when trained on large cohorts, WS-MIL methods tend to catch errors made by pathologists [19,98]. This emulates some current clinical workflows, in which pathologists examine cases concurrently and catch one another's misjudgments. It is especially important, as studies have shown that the diagnostic accuracy of pathologists depends to some extent on experience – junior pathologists tend to be more error-prone [114]. Thus, WS-MIL can help improve the skills of pathologists in training. Moreover, though it may be a hard truth to swallow for physicians, some pathologists are better than others with similar levels of experience [114]. Thus, WS-MIL can "raise the bar" of the average performance of pathologists across the board.

Whereas patient stratification reduces the magnitude of cases, WS-MIL enhanced pathology also has the potential to reduce clinician workload in terms of time spent on each case. The correct diagnosis for certain diseases requires pathologists to identify a few cells out of millions (i.e., "finding a needle in a haystack") [33,48]. Such tasks may take several hours for pathologists to make a single diagnosis. With WS-MIL assistance, relevant ROIs could be automatically identified in mere seconds. Then, pathologists would perform the relevant task, with a large percentage of the WSI automatically ignored. In other words, WS-MIL can help pathologists focus on relevant areas in the tissue. Some studies have already shown how WS-MIL methods can remove noise from H&E WSIs (i.e., remove tissue not correlated to outcome of interest) [48,99] and can accurately detect microanatomy and cellular mimics, which are sometimes confounded by pathologists [100]. Other studies have shown how WS-MIL can reduce labor-intensive and time-consuming (i.e., repetitive) tasks such as tumor labeling [41,49,94, 100], and Gleason grading [88]. Finally, one study concerning Fuhrman grading of CCRCC showed that exposing junior and senior pathologists to model-identified ROIs significantly improved their respective grading accuracy [89].

4.4. Personalized/precision medicine

Leveraging big data derived from clinical records and fusing said data from many components of patient care (i.e., multimodal data) would enable truly personalized medicine beyond the capabilities of any single clinician or team of clinicians. As previously discussed, WS-MIL models outperform all other common clinical or pathological features for predicting survival [95]. More to the point, several studies have demonstrated that *fusing* data across the patient care spectrum (i.e., clinical variables) with multiple WS-MIL methods for H&E WSI analysis more accurately predicts survival than any model alone and paves the way for personalized (i.e., patient-specific, data-driven) medicine [32–34,40]. Elsewhere, it has been discussed how next-generation sequencing (NGS) techniques (for histological subtyping, genetic mutations, molecular profiling, etc.) can be accurately modeled with WS-MIL on H&E [28,65] and allow for pinpointing the exact characteristics of a patient's cancer, enabling precision. In addition to the assays that inform treatments, WS-MIL methods can also estimate drug responses [48,98] and the efficacy of specific targeted therapies for patients [96,98,100] – for example, the efficacy of neoadjuvant chemoradiotherapy [49] – even in early-stage cancers which can benefit from more aggressive targeted treatment to prevent progression [80]. One study, in particular, showed that by combining WSIs from multiple sites, response to neoadjuvant chemotherapy response in triple-negative breast cancer could be more accurately predicted than through conventional clinical means [72]. All in all, from diagnosis to prognosis to treatment, WS-MIL methods stand to enable and enhance personalized medicine.

4.5. Clinical relevance and real-world scenarios

Histopathology contains rich sources of information beyond human perception that through WS-MIL enhanced histopathology can be leveraged to discover new imaging-based biomarkers for disease diagnosis, stratification, prognosis, and treatment. For diagnosis, several studies purport the hypothetical opportunities for WS-MIL enhanced histopathology to detect lesions and to screen, verify, or validate exploratory imaging biomarkers of tissue damage by pointing to lesions associated with the biomarkers, enabling phenotypic anchoring (which would otherwise be not possible or suboptimal) [28,67,90]. For example, the WS-MIL method by Lu et al. can differentiate between metastatic tumors and primary tumors as well as predict the origin of metastatic tumors, tasks that are challenging for pathologists [19]. Model-attended regions suggest that 1) 'dirty necrosis' and variably sized glands with densely packed, hyperchromatic nuclei, 2) sheets of cells as well as small tubules and glands, and 3) hyperchromatic nuclei and high nuclear to cytoplasmic ratios are imaging-based biomarkers for metastatic colorectal adenocarcinoma, breast carcinomas, and lung carcinomas, respectively. In a similar fashion, Brockmeller et al. linked inflamed fat in CRC with the presence of lymph node metastasis, which is supported somewhat by mechanistic studies [78].

For stratification, WS-MIL may tie genotypic markers to phenotypic markers – in other words, image-based morphological features related to

mutations, pathways, or overexpressed genes [28]. Some studies skirt the line of these opportunities. For example, histological ROIs on H&E that were associated with molecular biomarkers such as HER2 have been presented but not interpreted by pathologists [67,70]. Likewise, some discuss the opportunity for elucidating how molecular alterations drive the biological mechanisms that result in tumor growth yet give no evidence of the sort [67,68]. Few studies take that leap of faith forward and demonstratively tie genotypic markers to phenotypic markers. For example, Bilal et al. discuss how the automated analysis of the cellular composition of predictive histological features could improve our understanding of the downstream impact of these features and lead to new insights into representative and discriminative morphological features corresponding to molecular pathways and mutations for cancer [80]. They then showed a strong correlation between MSI and the infiltrate of inflammatory cells [80] as determined by their WS-MIL model. Kather et al. showed how poorly differentiated tumors were highlighted for CM1, well-differentiated glands for CMS2–3, and highly stromal ROIs for CMS4. Furthermore, they showed that FRAF mutations were associated with poor differentiation and mucinous areas, consistent with previous studies [68]. Finally, Chen et al. showed that in breast adenocarcinoma, genomic-guided WS-MIL for tumor suppression, protein kinases, and cellular differentiation generally reflected normal stroma, glands, and adipocytes [33].

Like with the diagnosis of disease, WS-MIL enhanced histopathology presents the opportunity to screen, verify, or validate *prognostic* imaging biomarkers of particular diseases [28,68,90]. For example, Saillard et al. demonstrate that a proangiogenic phenotype (identified by the model) was associated with poor clinical outcomes (previously confirmed in other research) [95]. Moreover, current prognostic markers can be quantified with WS-MIL – for example, quantifying molecular features from H&E (like immune infiltration, associated with poor prognosis) [28]. New prognostic imaging biomarkers can also be refined, such as gene signatures [40], or discovered and targeted in therapeutic treatments. – for example, stromal regions in mesothelioma (associated with good prognosis) [101], vascular spaces in hepatocellular carcinoma (associated with poor prognosis) [95], necrosis, dense tumor aggregates, desmoplastic stroma containing tumor infiltrates (associated with good prognosis) [32], and lymphocytes aggregates and normal stroma (associated with good prognosis) [32]. Finally, the prognosis may be directly estimated from H&E [48,98] or by fusing WS-MIL models with genomic and molecular info in multi-modal analyses, a task invariably impossible for humans [33,34,40].

Finally, some studies have proposed opportunities for WS-MIL on H&E and treatment. Xu et al. demonstrated via their WS-MIL model the intra-tumor variability of CIN. Based on their results, they suggested that therapeutics should focus on impacting high CIN tumor cells (in terms of their development and administration) [71]. Other studies discuss how treatment response may be predicted with WS-MIL [28,48,98] and that perhaps treatment decisions by oncologists should be guided by WS-MIL in the context of immunotherapy given the multi-faceted and multi-modal nature of cancer characterization [28]. Though several studies have predicted response to treatment for cancer done with full supervision [115], only one published study has done so using WS-MIL [72].

Ultimately, the opportunities discussed here are dependent on the viability of WS-MIL driven histopathology in oncology. Pathologists and likewise oncologists need to know why the WS-MIL systems they may use make the decisions they do *and* when said systems are uncertain. Explainability and interpretability are key aspects of this process of trust. It is no wonder then why WS-MIL has been so immensely popular for WS-MIL driven histopathology given its inherent interpretability (via visualization). Every study cited here identifies their models' interpretability as one of its defining features. For relatively simple tasks such as tumor detection, it is clear that WS-MIL models attend to tumor ROIs. This has led the to the clinical commercialization of Paige Prostate, a product that utilizes WS-MIL for H&E WSI (based on [4]). It was FDA-approved for clinical use to perform tumor detection on prostate biopsies. Paige also has several research-use-only products based on WS-MIL that will eventually find their way to a clinical market. Moreover, several other companies focusing on computational pathology – PathAI, Tempus Labs – are following suit, developing their own WS-MIL technologies. Eventually, given the ease with which they are explainable and interpretable, more WS-MIL methods will be commercialized, such as those for tumor subtyping or metastasis detection. However, the clinical utility and potential for automated metastasis detection as well as the prediction of molecular markers remains to be seen. Validation of these methods may require massive investment for commercialization and clinical use. For genetic markers, validation methods are still unavailable, thus, commercialization and clinical potential are probably unlikely in the near future. Ultimately, translation of these advanced WS-MIL methodologies to the clinic in a field that is quite heavily regulated will take quite a bit of time.

5. Challenges

Though clear opportunities do exist for the utilization of H&E slides to model clinical data, there still exist manifold challenges within the domain. Several recent review articles have discussed these challenges with respect to AI in histopathology [116], with some focusing on ethics (fairness [117,118], privacy [119,120], potential harm [118,119], equity [120]) and regulatory [12,26,118] issues. Here, we focus on challenges specific to WS-MIL. These include explainability and interpretability, validation via adjacent tissue sections, reproducibility, data availability, computational needs, data requirements, domain adaptability, external validation, dataset imbalances, and commercialization and clinical potential.

5.1. Explainability and interpretability

Though various WS-MIL driven methods verifiably attend to pathologically-relevant areas of H&E slides, they do not in some cases. Camelyon16 [16] serves as a case in point, consisting of sentinel lymph nodes that either present or absent breast cancer metastases. The main limitation of methods when applied to Camelyon16 (and perceptively deduced by Courtiol et al. as early as 2018 [48]) is that metastases are highly localized and restricted to a small area within the WSI. This is a non-issue for large metastases – models have been shown to attend to tumor regions [4,18,43–48,50]. However, when visualizing model-attended ROIs indicates for WSIs with small metastases (i.e. fewer than 10 tumor cells), the model *does not* attend to the metastases themselves, even when the WSI is classified correctly. It is no surprise then that no study qualitatively or quantitatively reports their model's performance on small metastases, despite Camelyon16 being the most popular dataset to apply WS-MIL. However, a few do *mention* that attention MIL attends to areas outside the tumor in Camelyon16 [43,46,74], leading to false negatives. As an additional note, several Gleason grading AI methods do not mimic the.

clinical process of individually grading each gland, so their interpretability is limited [88].

5.2. Validation of molecular marker-driven WS-MIL methods

WS-MIL methods virtualizing molecular assays such as gene expression and ODX will require corresponding IHC or ODX scores for model development and thus will require significant investment. For example, model-attended ROIs of DLBCL vs. non-DLBCL were confirmed on adjacent CD20 but would be expensive on a larger scale [41]. Similarly, weak supervision for HER2 scoring from H&E would need IHC to confirm and would be expensive or unavailable [70]. Finally, some studies would require spatial transcriptomics for validation, which is magnitudes more expensive [66]. Even if enough monetary resources were available for such targets, registration (i.e., digitally aligning

adjacent WSIs so that anatomies overlap) would be computationally infeasible for large datasets [70]. Moreover, the thickness of the tissue may obfuscate structures needed for registration [70]. Finally, the tools we have to measure signals of interest may not be sufficient. For example, there are assumptions about tissue heterogeneity (i.e., a molecular target from one tissue sample is the same in another sample) [66, 70]. Precise molecular annotations may alleviate this necessity [65]. Furthermore, some studies show that RNA-seq and gene expression are decorrelated, even though the former is used as a proxy for the latter [28]. Similarly, aneuploidy burden is used as a proxy for CIN, even though some studies show only a moderate correlation [71].

5.3. Reproducibility

Reproducibility refers to the ability of other researchers to reproduce the results of specific models on specific datasets utilizing specified protocols. It is to be expected that different studies applying the same methods (as comparisons) to the same datasets will report slightly different results, yet there is a great deal of variation that needs to be explored and explained. Tables 2 and 3 summarize the applications and reproductions of several methods to the same datasets. Clearly, there exists a problem reproducing past methods' results on the same datasets. This is particularly interesting for well-established methods, as their code is readily available and runs on the given datasets seamlessly. Also interesting, no study that was unable to reproduce previous models' results discussed *why* they were unable to do so. There are a few reasons why such variation may be observed. First, descriptions of methods are sometimes not given. For example, the simple pre-processing step of detecting the whitespace (non-tissue area) of WSIs is not described in some studies [33,44,45,50,83,104]. Second, method development tends to be an exhaustive process, whereas method reproduction (i.e., running someone else's code) is superficial. In other words, every effort is given to yield the best results for *my* method, but the bare minimum is given to reproduce *another's*. Third, there are no established experimental designs for the datasets mentioned. Typically, cross-validation is used, but several flavors exist, and implementations vary.

To further complicate matters, some studies do not reproduce previous methods and opt to directly report previous results [121]. Likewise, some do not compare their method to recent state-of-the-art methods. For example, Lu et al. [18] based their method on attention-based MIL [122] and did not provide a comparison to AB-MIL. Likewise, Wulczyn et al. [31] and Saillard et al. [95] did not provide any comparisons to past WS-MIL WSI survival prediction methods. And finally, Schumach et al. [28] provide no comparison methods for gene expression prediction. On a similar note, some studies conveniently do not utilize well-studied datasets for the task at hand and opt for other publicly available datasets. For example, Chen et al. in several studies [32,33,40] apply their methods to predict survival on five TCGA datasets and provide comparison methods even though three *other* TCGA datasets had been well-studied (see Table 3). Similarly, Shao et al. [92] utilize TCGA-LUSC as others in the past (see Table 3) but also TCGA-KIRC and TCGA-LIHC, unlike every past study. For these reasons, most state-of-the-art methods or well-established dataset and their results lack reproducibility. It is incumbent on reviewers, editors, and field leaders to develop standardized methods for presenting research findings in a systematic manner.

Table 2
Various methods for classifying Camelyon16, subtyping TCGA-NSCLC, and subtyping TCGA-RCC [46][0], [50][1], [47][2], [18][3], [4][4], [46][5], [43][6], [74][7], [55][8], [44][9], [61][10], [54][11], [99][12], [48][13] [51][14], [63][15], [57][16], [59][17], [53][18], [60][19], [56][20], [58][20], [63][21], [52], "NR" indicates that the original study did not report the result on the given dataset. "-" indicates that only one study performed the experiment, thus no comparison can be reported. Metrics are AUC. Methods are in chronological order. The bolded results are original studies.

		Method					
		DTFD-MIL[0]	TransMIL[1]	DS-MIL[2]	CLAM[3]	MIL-RNN[4]	AB-MIL
Dataset	Camelyon16	**0.946**[0]	**0.9309**[1]	**0.8944**[2]	**0.936**[3]	**0.899**[4]	NR
		0.9588[16]	0.906[5]	0.899[5]	0.858[9]	0.588[9]	0.854[5]
		0.946[18]	0.7748[6]	0.8179[1]	0.878[5]	0.875[5]	0.876[1]
		0.8836[20]	0.906[18]	0.9165[7]	0.8679[1]	0.888[1]	0.8653[2]
		0.932[21]	0.9259[20]	0.894[8]	0.884[8]	0.8064[7]	0.840[7]
			0.883[21]	0.8832[10]	0.8938[10]	0.806[8]	0.7504[6]
			0.928[22]	0.7544[11]	0.9059[17]	0.6913[11]	0.865[8]
				0.8539[14]	0.871[18]	0.8606[17]	0.8939[10]
				0.9070[17]	0.8580[19]	0.875[18]	0.6612[11]
				0.899[18]	0.9131[20]	0.861[21]	0.9105[17]
				0.892[21]	0.875[21]		0.854[18]
				0.875[22]	0.886[22]		0.851[21]
							0.840[22]
	TCGA-NSCLC	**0.961**[0]	**0.9603**[1]	**0.9633**[2]	**0.963**[3]	NR	NR
		0.9555[16]	0.949[5]	0.939[5]	0.949[5]	0.894[5]	0.941[5]
			0.9318[6]	0.8925[1]	0.9377[1]	0.9213[2]	0.9551[2]
			0.941[22]	0.859[10]	0.942[10]	0.9107[1]	0.8656[1]
				0.9633[11]	0.9037[12]	0.9168[12]	0.9465[6]
				0.920[15]	0.928[15]	0.9107[11]	0.8426[10]
				0.9461[17]	0.9520[17]	0.8178[17]	0.9488[11]
				0.924[22]	0.937[22]		0.8662[17]
							0.893[21]
							0.915[22]
	TCGA-RCC	-	**0.9882**[1]	NR	**0.991**[3]	NR	NR
			0.9826[6]	0.9841[1]	0.9799[1]	0.8831[17]	0.9702[1]
			0.965[22]	0.963[8]	0.979[8]		0.9778[6]
				0.971[15]	0.973[14]		0.923[8]
				0.9468[17]	0.9474[17]		0.9308[17]
				0.947[22]	0.967[22]		0.855[21]
							0.933[22]
	TCGA-BRCA	-	-	NR	-	-	NR
				0.838[15]			0.843[21]
				0.875[20]			0.869[20]

Table 3
Various methods for predicting survival [39][1], [113][2], [83][3], [92][4]. "NR" indicates that the original study did not report the result on the given dataset. "-" indicates that only one study performed the experiment, thus no comparison can be reported. C-index is a metric for predicting survival. Methods are in chronological order. The bolded results are original studies.

Method	Datasets		
	NLST	GBM	LUSC
DeepAttnMISL[3]	-	NR	NR
DeepGraphSurv[2]	-	-	**0.6606[2]**
			0.717[4]
WSISA-Lasso-Cox[1]	**0.703[1]**	**0.600[1]**	**0.638[1]**
	0.6380[2]	0.5760[2]	0.6380[2]
	0.5996[3]		
WSISA-MTLSA[1]	**0.680[1]**	-	-
	0.6305[3]		
Lasso-Cox	0.503[1]	0.440[1]	0.540[1]
	0.5280[2]	0.5574[2]	0.4738[2]
	0.4842[3]		0.527[4]
EnCox	0.502[1]	0.440[1]	0.613[1]
	0.4883[2]	0.5597[2]	0.4883[2]
			0.552[4]
Cox-Log	0.466[1]	-	-
	0.4998[3]		
Cox-Weibull	0.480[1]	-	-
	0.5577[3]		
RSF	0.485[1]	0.560[1]	0.347[1]
	0.5066[2]	0.5570[2]	0.5964[2]
			0.561[4]
BoostCI	0.511[1]	0.507[1]	0.339[1]
	0.5633[2]	0.5543[2]	0.5633[2]
	0.5595[3]		
MTLSA	0.609[1]	0.571[1]	0.536[1]
	0.5386[2]	0.5787[2]	0.5386[2]
	0.5053[3]		

5.4. Data availability

Scientific inquiry necessitates an open-source model for data sharing as a mechanism for peer-driven replication of studies, and though models and code are generally made publicly available, the in-house datasets from which they are derived are unavailable and thus make it impossible to reproduce their experiments. Notable methods include MIL-RNN [4], CLAM [18,19], and TOAD [40]. All provide code but not the internal datasets on which their models were trained. Such studies have also popularized an external validation method in which models are trained on internal datasets and then tested on publicly available datasets, such as TCGA or Camelyon16. From a model validation perspective, this is perfectly valid and in fact the gold standard for external validation – a dataset completely outside the institution from which the model was developed. However, from an AI method development standpoint, it is a nightmare, as subsequent technical method improvements or novel methods cannot be directly compared, as it may be the quality of the internal data rather than the robustness of the method itself that contributes to an improvement in performance. Many studies do not adopt this method of model validation and instead train and validate their methods as well as those previously on publicly available datasets (see Tables 2 and 3), although this is still wrought with issues (see the Section on Reproducibility).

The lack of an open-source model for data sharing also contributes to an over-reliance and artificial technical benchmarks, such as Camelyon16 and TCGA. This is exemplified in the current review – 28 of the articles train and validate their methods on just Camelyon16. Several dozen do the same using TCGA. Though benchmarks are a cornerstone for computer vision applications [123], they limit clinical viability, as the magnitude of the clinical applications addressed in WS-MIL is not proportional to the magnitude of the clinical problem in routine histopathological workflows. Very few research articles train and validate their methods on clinical datasets. However, without large, open-source clinical datasets, reliance on datasets such as Camelyon16 and TCGA will likely perpetuate, as there must be standard datasets to compare methods in computer vision.

5.5. Computational needs

At the most basic level, AI models need specialized hardware to be developed and deployed which may be prohibitively expensive in resource-constrained clinical settings. For example, one study utilized over 12,000 h of training time distributed over ten graphics-processing units (GPUs) to develop and validate their models [68]. On standard high-performance computing hardware, this likely would have taken a couple of orders of magnitude longer – on a standard desktop machine, several orders of magnitude longer. However, development and validation should not be confused with deployment. Deployment is much less costly in terms of computational needs. For example, two studies were able to deploy their model on a mobile device [68,115] to perform inference on WSIs.

At another level, several studies perform experiments utilizing hardware resources exclusive to a few heavily-invested institutions, making it impossible to reproduce results or utilize models. For example, Chuang et al. utilized Taiwania 2, a supercomputer equipped with 252 high-performance computing nodes, each with eight GPUs, totaling 256 GB of GPU RAM per node [75]. This is so much memory that it is the only study that has reported the ability to fit an entire WSI on a GPU. By comparison, the majority of studies presented here utilize GPUs with a median of 16 GB RAM and a maximum of 80 GB RAM, which necessitates splitting a WSI up into chunks to be processed. Under these conditions, it is impossible to reproduce Chuang et al.'s results or to utilize their model unless similar hardware were to be acquired.

5.6. Data requirements

Though the methods discussed here do not require annotations, it appears that WS-MIL has the effect of requiring (generally) larger amounts of data [4]. This is due to the inherent nature of weak labeling. Weak labels are noisy labels – they describe only one aspect of a WSI. For example, for cancer presence or absence, the task is to predict if a WSI contains tumor. A WSI with tumor also has normal tissue, meaning that the weak label of "tumor" is also being applied to all the normal tissue. What WS-MIL aims to do is to learn to suppress that noisy, label-irrelevant information while bolstering relevant (i.e., discriminative) ROIs. However, to learn the true signal of a tumor, many examples (large datasets) are needed. Unfortunately, this reliance on larger datasets ultimately limits the generalizability of studies utilizing smaller datasets [4]. Yet, some studies have successfully applied WS-MIL to H&E WSIs using relatively small datasets. For example, Lu et al. report that their method is data-efficient (i.e., not requiring many slides) [18] and therefore does not require as many slides as previous studies [4]. Likewise, some researchers focus specifically on developing WS-MIL methods that do not require many WSIs during training [88].

5.7. Domain adaptability

Unlike molecular assays such as gene expression panels and spatial transcriptomics which are agnostic to different diseases and organ types, WS-MIL driven histopathology methods are limited to the disease and tissue type that they were developed on. Most models only examine one disease, tissue type, and outcome combination [4,65,82]. Perhaps the best example of this limitation comes from Kather et al. Their incredibly large-scale study examining 14 cancers and predicting the mutation status of 95 genes, 20 gene signatures, and 17 standard-of-care variables from H&E WSIs trained a separate model for each combination cancer/tissue type/outcome combination, resulting in 1848 distinct WS-MIL models [68]. Such a paradigm would be infeasible in a practical clinical setting. A WS-MIL model cannot be trained for every scenario.

Some hope may be offered by the fact that low-level features (i.e., edges, points, blobs, textures) are generalizable across domains in natural images and so the same is probably true for histopathology [90]. This is exemplified by one study in particular by Lu et al. [19] in which they trained *one model* to predict the site-of-origin for cancers of unknown primary across 18 origins of cancer. Another path forward to avoid granularity among WS-MIL models is something known as open-set learning [124]. In this paradigm, WS-MIL models can predict "I don't know." This would be especially important for corner cases that WS-MIL models may never see during development but may nonetheless appear during deployment. Such cases could be deferred to experts.

5.8. Homogeneity of data

Generalizability refers to how well a model performs on out-of-distribution data (i.e., from a different institution, digitization protocol, hardware, etc.). The overwhelming majority of WS-MIL methods applied to H&E WSIs are validated using data generated from the same institution, scanner, and so on. Even large-scale studies such as that by Butlen et al. with 5759 H&E WSIs [15] or Chuang et al. with 3182 H&E WSIs [75] tested their methods on an internal dataset. Furthermore, several studies continually train and validate novel methods to TCGA [18,43,46,47,50,54,96,98,99] with no external validation knowing full well that TCGA may not fully represent the diversity and heterogeneity of tissues that pathologists inspect [98]. Finally, some studies train and validate exclusively on frozen sections, which tends to lead to better performance than formalin-fixed, paraffin-embedded (FFPE) despite the fact that frozen sections only constitute a small fraction of diagnostic WSIs [68]. This is in stark contrast to the gold standard of model validation that applies models to external FFPE WSIs (i.e., a different institution, scanner, etc.). Par the course, some studies make it a point to gather data from multiple institutions, such as Lu et al. [19] which collected WSIs from 223 institutions. As mentioned before, several studies externally validate their methods only with publicly available datasets [4,18,40]. Ideally, all studies should adopt this external validation paradigm, but as discussed before (see data availability), this comes with its own issues.

5.9. Sampling issues

The robustness with which WS-MIL learns the structure and variation of tissues and their diseased pathologies is inversely proportional to the amount of that healthy or diseased tissue there is within the dataset. The same can be said generally of WS-MIL – the less it sees, the less it understands. In the context of oncology, this can manifest in several manners. For example (and as discussed before), corner cases. Corner cases represent the bane of WS-MIL's existence. They are anything so rare that any given dataset (or patient cohort in the context of oncology) fails to represent it adequately or at all. Therefore, WS-MIL methods do not learn how to identify said corner cases. For histopathology in particular, it is especially difficult to capture the immense diversity and heterogeneity present in histology for rare diagnoses [18]. This goes even beyond the disease-level for histopathology. Indeed, the same "corner case" problem can be analogized to rarity in terms of the magnitude of tissue representation. Case in point are metastases, specifically micrometastases. They physically represent an extremely small (i.e., rare) portion of a WSI. Therefore, it is difficult for WS-MIL methods to learn what micrometastases look like especially using WS-MIL, and therefore accurately detect them [18]. These problems are compounded by the redundancy in WSIs (i.e., repetitive tissue structures) and noise (i.e., tissue structures unrelated to the weak label), although some studies acknowledge and try to resolve these problems [46,66], one in particular with remarkable successes [51].

5.10. Commercialization and clinical potential

Ultimately, the shortcomings discussed here, especially explainability and interpretability, limit the viability of WS-MIL driven histopathology in oncology. Pathologists and likewise oncologists need to know why the WS-MIL systems they may use make the decisions they do *and* when said systems are uncertain. Explainability and interpretability are key aspects of this process of trust. It is no wonder then why WS-MIL has been so immensely popular for WS-MIL driven histopathology given its inherent interpretability (via visualization). Every study cited here identifies their models' interpretability as one of its defining features. For relatively simple tasks such as tumor detection, it is clear that WS-MIL models attend to tumor ROIs. That is why the only commercial product that utilizes WS-MIL for H&E WSI (based on [4]) and has been FDA-approved for clinical use performs tumor detection. Eventually, given the ease with which they are explainable and interpretable, more WS-MIL methods will be commercialized, such as those for tumor subtyping or metastasis detection. However, the clinical potential for automated metastasis detection as well as the prediction of molecular markers remains to be seen. Validation of these methods may require massive investment for commercialization and clinical use. For genetic markers, validation methods are still unavailable, thus, commercialization and clinical potential are probably unlikely in the near future.

6. Summary

WS-MIL enhanced histopathology presents unprecedented opportunities to benefit oncology through WS-MIL methods that require only patient-level clinical data (i.e., weak labels) and respective H&E slides. To support this area of research, we have presented a summary of various applications of WS-MIL methods to H&E WSIs with no annotations. We divided these methods first based on their degree of verifiability and then by commonly recurring application areas.

First, we discussed methods to predict morphological markers of disease such as tumor presence or absence, metastases, subtypes, and grades. These methods were entirely interpretable and verifiable given that model-identified ROIs could be overlapped with pathologist-annotated ROIs. Second, we discussed methods to predict molecular markers of disease such as gene expression and molecular subtyping which cannot be verified with pathologist ROIs but could be visualized and then validated (albeit with moderate difficulty) based on overlap with adjacent IHC, immunofluorescence, or digital spatial profiling. Third, we discussed methods to predict genetic markers of disease such as mutations, mutational burden, MSI, CIN, and HRD. Unlike the previous applications, these models cannot be validated with current technologies and would require that specific genetic alterations can be spatially resolved to confirm if model-attended regions truly correspond to tumor cells with said alterations. Fourth, we discussed direct prediction of survival and how (similar to genetic markers) there are histopathological features that correlate but nonetheless cannot be tied directly to specific survival times and are thus unverifiable.

Finally, we discussed the opportunities and challenges for WS-MIL applied to H&E WSIs. Opportunities included reducing the cost of research through a one-label-per-slide paradigm and through reducing annotation burden; reducing the cost of clinical care through virtualization of specialized assays; reducing the workload of clinicians through triaging, offering second opinions, and automatic processing of repetitive and time-consuming tasks; personalizing medicine by leveraging multi-modal data beyond the capabilities of any one clinician or team; and unlocking the full potential of histopathology through discovery of new imaging-based biomarkers for disease diagnosis, stratification, prognosis, and treatment. Current challenges included the limitations of explainability and interpretability, especially in the case of micrometastases; the cost and technical limitations of validating molecular marker-driven AI methods; issues with reproducing past methods' results on the same datasets; making large internal datasets available for

peer-driven replication; specialized computational limits posed in resource-limited environments as well as unmatchable computational resources prohibiting replication; the data requirements imposed by WS-MIL; the lack of adaptability of models trained on specific diseases and outcomes; the apparent lack of external validation; patient-level and tissue-level corner cases and imbalances; and finally limitations imposed on commercialization and clinical potential based on these current challenges. Ultimately, the relative ease and minimum upfront cost with which relevant data can be collected in addition to the plethora of available WS-MIL methods for H&E WSI outcome-driven analysis will surmount these current limitations and achieve the innumerable opportunities associated with AI-driven histopathology for the benefit of oncology.

6.1. Search strategies and selection criteria

We used Google Scholar and PubMed to find relevant manuscripts. We restricted our search to papers published in English between Jan 1, 2018 and July 15, 2023. We used the following terms in different combinations: “weak”, “weak labels”, “weak supervision”, “weakly supervised”, “slide-level”, “slide-level labels”, “whole slide”, “whole slide image”, “whole slide imaging”, “WSI”, “multiple instance learning”, “MIL”, “histopathology”, “deep learning”, “machine learning”, “artificial intelligence”, “AI”, “digital pathology”, “computational pathology”, “subtyping”, “grading”, “gene expression”, “RNAseq”, “molecular subtyping”, “genetic markers”, “microsatellite stability”, “MSI”, “mutational burden”, “chromosomal instability”, “CIN”, “HRD”, “survival”, “survival prediction”, “imaging biomarkers”. We included studies that utilized no pixel-level labels or pre-trained histopathology models (that were trained with pixel labels) and tried to predict slide-level labels. Studies that utilized pixel-level labels for verification were allowed. Finally, we used the “Cited by” function on Google Scholar to find additional studies suiting our criteria.

Funding Source

The research was partially funded by a National Institutes of Health R01CA276301 (PIs: Niazi, Wei), Trailblazer award R21EB029493 (PIs: Niazi, Segal), R21CA273665 (PI: Gurcan), R01DC020715 (PIs: Gurcan and Moberly), and Alliance Clinical Trials in Oncology GR125886 (PIs: Frankel and Niazi). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or the Alliance Clinical Trials in Oncology.

Declaration of Competing Interest

The authors declared that they have no conflicts of interest in this work.

Data availability

No data was used for the research described in the article.

References

[1] M.K.K. Niazi, A.V. Parwani, M.N. Gurcan, Digital pathology and artificial intelligence, Lancet Oncol. 20 (5) (2019) e253–e261.
[2] S. Sornapudi, et al., Deep learning nuclei detection in digitized histology images by superpixels, J. Pathol. Inform. 9 (2018).
[3] C. Li, et al., DeepMitosis: Mitosis detection via deep detection, verification and segmentation networks, Med. Image Anal. 45 (2018) 121–133.
[4] G. Campanella, et al., Clinical-grade computational pathology using weakly supervised deep learning on whole slide images, Nat. Med. 25 (8) (2019) 1301–1309.
[5] E. Abels, et al., Computational pathology definitions, best practices, and recommendations for regulatory guidance: a white paper from the Digital Pathology Association, J. Pathol. 249 (3) (2019) 286–294.
[6] N. Kumar, et al., A dataset and a technique for generalized nuclear segmentation for computational pathology, IEEE Trans. Med. Imaging 36 (7) (2017) 1550–1560.
[7] R. Verma, et al., MoNuSAC2020: A multi-organ nuclei segmentation and classification challenge, IEEE Trans. Med. Imaging 40 (12) (2021) 3413–3423.
[8] K. Sirinukunwattana, et al., Gland segmentation in colon histology images: The glas challenge contest, Med. Image Anal. 35 (2017) 489–502.
[9] S. Graham, et al., Hover-net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images, Med. Image Anal. 58 (2019), 101563.
[10] Z. Swiderska-Chadaj, et al., Learning to detect lymphocytes in immunohistochemistry with deep learning, Med. Image Anal. 58 (2019), 101547.
[11] V. Baxi, et al., Digital pathology and artificial intelligence in translational medicine and clinical practice, Mod. Pathol. 35 (1) (2022) 23–32.
[12] A. Echle, et al., Deep learning in cancer pathology: a new generation of clinical biomarkers, Br. J. Cancer 124 (4) (2021) 686–696.
[13] Y. Jiang, et al., Emerging role of deep learning-based artificial intelligence in tumor pathology, Cancer Commun. 40 (4) (2020) 154–166.
[14] T.E. Tavolara, et al., Automatic generation of the ground truth for tumor budding using H&E stained slides. Medical Imaging 2022: Digital and Computational Pathology, SPIE,, 2022.
[15] W. Bulten, et al., Automated deep-learning system for Gleason grading of prostate cancer using biopsies: a diagnostic study, Lancet Oncol. 21 (2) (2020) 233–241.
[16] B.E. Bejnordi, et al., Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer, Jama 318 (22) (2017) 2199–2210.
[17] A. Bychkov, M. Schubert, Constant demand, patchy supply. The Pathologist, Texere,, 2023, pp. 18–27.
[18] M.Y. Lu, et al., Data-efficient and weakly supervised computational pathology on whole-slide images, Nat. Biomed. Eng. 5 (6) (2021) 555–570.
[19] M.Y. Lu, et al., AI-based pathology predicts origins for cancers of unknown primary, Nature 594 (7861) (2021) 106–110.
[20] M.Y. Lu, et al., Federated learning for computational pathology on gigapixel whole slide images, Med. Image Anal. 76 (2022), 102298.
[21] M.K.K. Niazi, et al., Relationship between the Ki67 index and its area based approximation in breast cancer, BMC Cancer 18 (1) (2018) 1–9.
[22] T.E. Tavolara, et al., A modular cGAN classification framework: Application to colorectal tumor detection, Sci. Rep. 9 (1) (2019) 1–8.
[23] O. Maron, T. Lozano-Pérez, A framework for multiple-instance learning, Adv. Neural Inf. Process. Syst. 10 (1997).
[24] M.M. Dundar, et al., A multiple instance learning approach toward optimal classification of pathology slides. International Conference on Pattern Recognition, IEEE,, 2010.
[25] M.M. Dundar, et al., Computerized classification of intraductal breast lesions using histopathological images, IEEE Trans. Biomed. Eng. 58 (7) (2011) 1977–1984.
[26] K. Bera, et al., Artificial intelligence in digital pathology—new tools for diagnosis and precision oncology, Nat. Rev. Clin. Oncol. 16 (11) (2019) 703–715.
[27] A. Kleppe, et al., Designing deep learning studies in cancer diagnostics, Nat. Rev. Cancer 21 (3) (2021) 199–211.
[28] B. Schmauch, et al., A deep learning model to predict RNA-Seq expression of tumours from whole slide images, Nat. Commun. 11 (1) (2020) 1–15.
[29] Y. Fu, et al., Pan-cancer computational histopathology reveals mutations, tumor composition and prognosis, Nat. Cancer 1 (8) (2020) 800–810.
[30] Q. Zheng, et al., Predicting Lymph Node Metastasis Status from Primary Muscle-Invasive Bladder Cancer Histology Slides Using Deep Learning: A Retrospective Multicenter Study, Cancers 15 (11) (2023) 3000.
[31] E. Wulczyn, et al., Deep learning-based survival prediction for multiple cancer types using histopathology images, PloS One 15 (6) (2020), e0233678.
[32] R.J. Chen, et al., Whole slide images are 2d point clouds: Context-aware survival prediction using patch-based graph convolutional networks. Medical Image Computing and Computer Assisted Intervention, Springer,, 2021.
[33] Chen, R.J., et al. *Multimodal co-attention transformer for survival prediction in gigapixel whole slide images*. in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021.
[34] R.J. Chen, et al., Pan-cancer integrative histology-genomic analysis via multimodal deep learning, Cancer Cell 40 (8) (2022) 865–878.
[35] M. Lu, et al., Smile: Sparse-attention based multiple instance contrastive learning for glioma sub-type classification using pathological images. MICCAI Workshop on Computational Pathology, PMLR,, 2021.
[36] L. Hou, et al., *Patch-based convolutional neural network for whole slide tissue image classification*, Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (2016).
[37] Z. Li, et al., Vision transformer-based weakly supervised histopathological image analysis of primary brain tumors, IScience 26 (1) (2023).
[38] O.L. Saldanha, et al., Self-supervised attention-based deep learning for pan-cancer mutation prediction from histopathology, NPJ Precis. Oncol. 7 (1) (2023) 35.
[39] Zhu, X., et al. Wsisa: Making survival prediction from whole slide histopathological images. in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017.
[40] Chen, R.J., et al., Pathomic fusion: an integrated framework for fusing histopathology and genomic features for cancer diagnosis and prognosis. IEEE Transactions on Medical Imaging, 2020.
[41] Hashimoto, N., et al. Multi-scale domain-adversarial multiple-instance CNN for cancer subtype classification with unannotated histopathological images. in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020.

[42] H. Xiang, et al., Multi-scale representation attention based deep multiple instance learning for gigapixel whole slide image analysis, Med. Image Anal. (2023), 102890.
[43] Javed, S.A., et al., Additive MIL: Intrinsic Interpretability for Pathology. arXiv preprint arXiv:2206.01794, 2022.
[44] Z. Su, et al., Attention2majority: Weak multiple instance learning for regenerative kidney grading on whole slide images, Med. Image Anal. 79 (2022), 102462.
[45] Y. Sharma, et al., *Cluster-to-conquer: A framework for end-to-end multi-instance learning for whole slide image classification.* Medical Imaging with Deep Learning, PMLR,, 2021.
[46] Zhang, H., et al. *DTFD-MIL*: Double-Tier Feature Distillation Multiple Instance Learning for Histopathology Whole Slide Image Classification. in Proceedings of the IEEE/CVF Conference on Computer Vision and *Pattern Recognition,* 2022.
[47] Li, B., Y. Li, and K.W. Eliceiri, Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021.
[48] Courtiol, P., et al., Classification and disease localization in histopathology using only global labels: A weakly-supervised approach. arXiv preprint arXiv: 1802.02212, 2018.
[49] D. Zhang, et al., Using multi-scale convolutional neural network based on multi-instance learning to predict the efficacy of neoadjuvant chemoradiotherapy for rectal cancer, IEEE J. Transl. Eng. Health Med. 10 (2022) 1–8.
[50] Z. Shao, et al., Transmil: Transformer based correlated multiple instance learning for whole slide image classification, Adv. Neural Inf. Process. Syst. 34 (2021) 2136–2147.
[51] Liu, K., et al., Multiple instance learning via iterative self-paced supervised contrastive learning. in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023.
[52] J.-G. Yu, et al., Bayesian collaborative learning for whole-slide image classification, IEEE Trans. Med. Imaging (2023).
[53] M.U. Oner, et al., Distribution based MIL pooling filters: Experiments on a lymph node metastases dataset, Med. Image Anal. 87 (2023), 102813.
[54] L. Qu, et al., *DGMIL: Distribution Guided Multiple Instance Learning for Whole Slide Image Classification.* International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer,, 2022.
[55] B. Shi, X. Liu, F. Zhang, MLCN: metric learning constrained network for whole slide image classification with bilinear gated attention mechanism. International Workshop on Computational Mathematics Modeling in Cancer Analysis, Springer,, 2022.
[56] S. Dooper, et al., Gigapixel end-to-end training using streaming and attention, Med. Image Anal. (2023), 102881.
[57] Lin, T., et al., Interventional bag multi-instance learning on whole-slide pathological images. in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023.
[58] Li, H., et al., Task-specific fine-tuning via variational information bottleneck for weakly-supervised pathology whole slide image classification. in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023.
[59] Z. Zhu, et al., Murcl: Multi-instance reinforcement contrastive learning for whole slide image classification, IEEE Trans. Med. Imaging (2022).
[60] U. Sajjad, et al., NRK-ABMIL: subtle metastatic deposits detection for predicting lymph node metastasis in breast cancer whole-slide images, Cancers 15 (13) (2023) 3428.
[61] W. Wu, et al., Clustering-based multi-instance learning network for whole slide image classification. International Workshop on Computational Mathematics Modeling in Cancer Analysis, Springer,, 2022.
[62] Lazard, T., et al., Giga-SSL: Self-Supervised Learning for Gigapixel Images. in Proceedings of the IEEE/CVF Conference on Computer Vision and *Pattern Recognition,* 2023.
[63] Chen, R.J., et al., Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022.
[64] Lu, M.Y., et al., Visual Language Pretrained Multiple Instance Zero-Shot Transfer for Histopathology Images. in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023.
[65] H. Qu, et al., Genetic mutation and biological pathway prediction based on whole slide images in breast carcinoma using deep learning, NPJ Precis. Oncol. 5 (1) (2021) 1–11.
[66] Weitz, P., et al., An investigation of attention mechanisms in histopathology whole-slide-image analysis for regression objectives. in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021.
[67] N. Naik, et al., Deep learning-enabled breast cancer hormonal receptor status determination from base-level H&E stains, Nat. Commun. 11 (1) (2020) 1–8.
[68] J.N. Kather, et al., Pan-cancer image-based detection of clinically actionable genetic alterations, Nat. Cancer 1 (8) (2020) 789–799.
[69] S.P. Oliveira, et al., Weakly-supervised classification of HER2 expression in breast cancer haematoxylin and eosin stained slides, Appl. Sci. 10 (14) (2020) 4728.
[70] T.E. Tavolara, et al., Predicting HER2 scores from registered HER2 and H&E images. Medical Imaging 2022: Digital and Computational Pathology, SPIE,, 2022.
[71] Z. Xu, et al., Deep learning predicts chromosomal instability from histopathology images, IScience 24 (5) (2021), 102394.
[72] J. Ogier du Terrail, et al., Federated learning for predicting histological response to neoadjuvant chemotherapy in triple-negative breast cancer, Nat. Med. 29 (1) (2023) 135–146.
[73] P.C. Neto, et al., iMIL4PATH: A Semi-Supervised Interpretable Approach for Colorectal Whole-Slide Images, Cancers 14 (10) (2022) 2489.
[74] N. Marini, et al., Multi-scale task multiple instance learning for the classification of digital pathology images with global annotations. MICCAI Workshop on Computational Pathology, PMLR,, 2021.
[75] W.-Y. Chuang, et al., Identification of nodal micrometastasis in colorectal cancer using deep learning on annotation-free whole-slide images, Mod. Pathol. 34 (10) (2021) 1901–1911.
[76] C. Zhou, et al., Histopathology classification and localization of colorectal cancer using global labels by weakly supervised deep learning, Comput. Med. Imaging Graph. 88 (2021), 101861.
[77] L. Tan, et al., Colorectal cancer lymph node metastasis prediction with weakly supervised transformer-based multi-instance learning, Med. Biol. Eng. Comput. 61 (6) (2023) 1565–1580.
[78] S. Brockmoeller, et al., Deep learning identifies inflamed fat as a risk factor for lymph node metastasis in early colorectal cancer, J. Pathol. 256 (3) (2022) 269–281.
[79] J.M. Niehues, et al., Generalizable biomarker prediction from cancer pathology slides with self-supervised deep learning: A retrospective multi-centric study, Cell Rep. Med. 4 (4) (2023).
[80] M. Bilal, 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 3 (12) (2021) e763–e772.
[81] Y. Schirris, et al., DeepSMILE: contrastive self-supervised pre-training benefits MSI and HRD classification directly from H&E whole-slide images in colorectal and breast cancer, Med. Image Anal. 79 (2022), 102464.
[82] P.L. Schrammen, et al., Weakly supervised annotation-free cancer detection and prediction of genotype in routine histopathology, J. Pathol. 256 (1) (2022) 50–60.
[83] J. Yao, et al., Whole slide images based cancer survival prediction using attention guided deep multiple instance learning networks, Med. Image Anal. 65 (2020), 101789.
[84] Y. Zheng, et al., Kernel Attention Transformer for Histopathology Whole Slide Image Analysis and Assistant Cancer Diagnosis, IEEE Trans. Med. Imaging (2023).
[85] S. Fremond, et al., Interpretable deep learning model to predict the molecular classification of endometrial cancer from haematoxylin and eosin-stained whole-slide images: a combined analysis of the PORTEC randomised trials and clinical cohorts, Lancet Digit. Health 5 (2) (2023) e71–e82.
[86] M.S. Wibawa, et al., Multi-scale attention-based multiple instance learning for classification of multi-gigapixel histology images. European Conference on Computer Vision, Springer,, 2022.
[87] T.E. Tavolara, et al., Grading and localization of histological features for bioengineered kidney constructs. Medical Imaging 2021: Digital Pathology, SPIE,, 2021.
[88] J. Silva-Rodríguez, et al., Self-learning for weakly supervised gleason grading of local patterns, IEEE J. Biomed. Health Inform. 25 (8) (2021) 3094–3104.
[89] Q. Zheng, et al., A Weakly Supervised Deep Learning Model and Human–Machine Fusion for Accurate Grading of Renal Cell Carcinoma from Histopathology Slides, Cancers 15 (12) (2023) 3198.
[90] C.A.C. Freyre, et al., Biomarker-based classification and localization of renal lesions using learned representations of histology—a machine learning approach to histopathology, Toxicol. Pathol. 49 (4) (2021) 798–814.
[91] A. Alsaafin, et al., Learning to predict RNA sequence expressions from whole slide images with applications for search and classification, Commun. Biol. 6 (1) (2023) 304.
[92] W. Shao, et al., Weakly supervised deep ordinal cox model for survival prediction from whole-slide pathological images, IEEE Trans. Med. Imaging 40 (12) (2021) 3739–3747.
[93] C. Sun, et al., Deep learning-based classification of liver cancer histopathology images using only global labels, IEEE J. Biomed. Health Inform. 24 (6) (2019) 1643–1651.
[94] H. Liao, et al., Deep learning-based classification and mutation prediction from histopathological images of hepatocellular carcinoma, Clin. Transl. Med. 10 (2) (2020).
[95] C. Saillard, et al., Predicting survival after hepatocellular carcinoma resection using deep learning on histological slides, Hepatology 72 (6) (2020) 2000–2013.
[96] F. Kanavati, et al., Weakly-supervised learning for lung carcinoma classification using deep learning, Sci. Rep. 10 (1) (2020) 1–11.
[97] L. Cao, et al., E2EFP-MIL: End-to-end and high-generalizability weakly supervised deep convolutional network for lung cancer classification from whole slide image, Med. Image Anal. 88 (2023), 102837.
[98] N. Coudray, et al., Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning, Nat. Med. 24 (10) (2018) 1559–1567.
[99] L. Zhao, et al., Lung cancer subtype classification using histopathological images based on weakly supervised multi-instance learning, Phys. Med. Biol. 66 (23) (2021), 235013.
[100] X. Wang, et al., Weakly supervised deep learning for whole slide lung cancer image analysis, IEEE Trans. Cybern. 50 (9) (2019) 3950–3962.
[101] P. Courtiol, et al., Deep learning-based classification of mesothelioma improves prediction of patient outcome, Nat. Med. 25 (10) (2019) 1519–1525.
[102] H. Xu, S. Park, T.H. Hwang, Computerized classification of prostate cancer gleason scores from whole slide images, IEEE/ACM Trans. Comput. Biol. Bioinforma. 17 (6) (2019) 1871–1882.

[103] Z. Yang, et al., The devil is in the details: a small-lesion sensitive weakly supervised learning framework for prostate cancer detection and grading, Virchows Arch. 482 (3) (2023) 525–538.
[104] D. Dov, et al., Weakly supervised instance learning for thyroid malignancy prediction from whole slide cytopathology images, Med. Image Anal. 67 (2021), 101814.
[105] D. Cifci, S. Foersch, J.N. Kather, Artificial intelligence to identify genetic alterations in conventional histopathology, J. Pathol. 257 (4) (2022) 430–444.
[106] A. Shmatko, et al., Artificial intelligence in histopathology: enhancing cancer research and clinical oncology, Nat. Cancer 3 (9) (2022) 1026–1038.
[107] L. Qu, M. Wang, Z. Song, Bi-directional weakly supervised knowledge distillation for whole slide image classification, Adv. Neural Inf. Process. Syst. 35 (2022) 15368–15381.
[108] P. Tourniaire, et al., Attention-based multiple instance learning with mixed supervision on the Camelyon16 dataset. MICCAI Workshop on Computational Pathology, PMLR,, 2021.
[109] J.N. Weinstein, et al., The cancer genome atlas pan-cancer analysis project, Nat. Genet. 45 (10) (2013) 1113–1120.
[110] P.A. Humphrey, Gleason grading and prognostic factors in carcinoma of the prostate, Mod. Pathol. 17 (3) (2004) 292–306.
[111] H. Pinckaers, et al., Detection of prostate cancer in whole-slide images through end-to-end training with image-level labels, IEEE Trans. Med. Imaging 40 (7) (2021) 1817–1826.
[112] T. Qaiser, et al., Her 2 challenge contest: a detailed assessment of automated her 2 scoring algorithms in whole slide images of breast cancer tissues, Histopathology 72 (2) (2018) 227–238.
[113] R. Li, et al., Graph CNN for survival analysis on whole slide pathological images. International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer,, 2018.
[114] A.A. Renshaw, E.W. Gould, Measuring errors in surgical pathology in real-life practice: defining what does and does not matter, Am. J. Clin. Pathol. 127 (1) (2007) 144–152.
[115] L.A. Hildebrand, et al., Artificial intelligence for histology-based detection of microsatellite instability and prediction of response to immunotherapy in colorectal cancer, Cancers 13 (3) (2021) 391.
[116] M.A. Berbís, et al., Computational pathology in 2030: a Delphi study forecasting the role of AI in pathology within the next decade, EBioMedicine 88 (2023).
[117] R.J. Chen, et al., Algorithmic fairness in artificial intelligence for medicine and healthcare, Nat. Biomed. Eng. 7 (6) (2023) 719–742.
[118] C. Chauhan, R.R. Gullapalli, Ethics of AI in pathology: current paradigms and emerging issues, Am. J. Pathol. 191 (10) (2021) 1673–1683.
[119] B.R. Jackson, et al., The ethics of artificial intelligence in pathology and laboratory medicine: principles and practice, in: Academic Pathology, 8, 2021, 2374289521990784.
[120] F. McKay, et al., The ethical challenges of artificial intelligence-driven digital pathology, J. Pathol.: Clin. Res. 8 (3) (2022) 209–216.
[121] P. Chikontwe, et al., Feature re-calibration based multiple instance learning for whole slide image classification. International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer,, 2022.
[122] M. Ilse, J. Tomczak, M. Welling, Attention-based deep multiple instance learning. International conference on machine learning, PMLR,, 2018.
[123] N.G. Laleh, et al., Benchmarking weakly-supervised deep learning pipelines for whole slide classification in computational pathology, Med. Image Anal. 79 (2022), 102474.
[124] W.J. Scheirer, et al., Toward open set recognition, IEEE Trans. Pattern Anal. Mach. Intell. 35 (7) (2012) 1757–1772.