Improving Workflow Integration with xPath: Design and Evaluation of a Human-AI Diagnosis System in Pathology

HONGYAN GU*, YUAN LIANG, and YIFAN XU, University of California, Los Angeles, USA
CHRISTOPHER KAZU WILLIAMS, SHINO MAGAKI, NEGAR KHANLOU, HARRY VINTERS, and ZESHENG CHEN, UCLA David Geffen School of Medicine, USA
SHUO NI, University of California, Los Angeles and University of Southern California, USA
CHUNXU YANG, University of California, Los Angeles and Peking University, USA
WENZHONG YAN, University of California, Los Angeles, USA
XINHAI ROBERT ZHANG, University of Texas Health Science Center at Houston, USA
YANG LI, Google Research, USA
MOHAMMAD HAERI*, University of Kansas Medical Center, USA
XIANG 'ANTHONY' CHEN, University of California, Los Angeles, USA

Recent developments in AI have provided assisting tools to support pathologists' diagnoses. However, it remains challenging to incorporate such tools into pathologists' practice; one main concern is AI's insufficient workflow integration with medical decisions. We observed pathologists' examination and discovered that the main hindering factor to integrate AI is its incompatibility with pathologists' workflow. To bridge the gap between pathologists and AI, we developed a human-AI collaborative diagnosis tool — xPath— that shares a similar examination process to that of pathologists, which can improve AI's integration into their routine examination. The viability of xPath is confirmed by a technical evaluation and work sessions with twelve medical professionals in pathology. This work identifies and addresses the challenge of incorporating AI models into pathology, which can offer first-hand knowledge about how HCI researchers can work with medical professionals side-by-side to bring technological advances to medical tasks towards practical applications.

CCS Concepts: • **Human-centered computing → Human computer interaction (HCI)**; • **Applied computing → Life and medical sciences**; • **Computing methodologies** → *Machine learning*.

Additional Key Words and Phrases: Human-AI collaboration; digital pathology; medical AI; meningioma

*Corresponding authors.

Authors' addresses: Hongyan Gu, ghy@ucla.edu; Yuan Liang, liangyuandg@ucla.edu; Yifan Xu, yifanxu@ucla.edu, 580 Portola Plaza, Room 1538, University of California, Los Angeles, Los Angeles, California, 90095, USA; Christopher Kazu Williams, ckwilliams@mednet.ucla.edu; Shino Magaki, smagaki@mednet.ucla.edu; Negar Khanlou, nkhanlou@mednet.ucla.edu; Harry Vinters, hvinters@mednet.ucla.edu; Zesheng Chen, zesheng.chen.med1@ssss.gouv.qc.ca, UCLA David Geffen School of Medicine, Los Angeles, California, USA, 90095; Shuo Ni, shuoni@usc.edu, University of California, Los Angeles and University of Southern California, Los Angeles, California, 90095, USA; Chunxu Yang, chunxuyang@ucla.edu, University of California, Los Angeles and Peking University, Los Angeles, California, 90095, USA; Wenzhong Yan, wzyan24@ucla.edu, University of California, Los Angeles, Los Angeles, California, 90095, USA; Xinhai Robert Zhang, Xinhai.R.Zhang@uth.tmc.edu, University of Texas Health Science Center at Houston, Houston, Texas, 77030, USA; Yang Li, yangli@acm.org, Google Research, Mountain View, California, 94043, USA; Mohammad Haeri, mhaeri@kumc.edu, University of Kansas Medical Center, Kansas City, Kansas, 66160, USA; Xiang 'Anthony' Chen, xac@ucla.edu, University of California, Los Angeles, Los Angeles, California, 90095, USA.

1073-0516/2022/12-ART
https://doi.org/10.1145/3577011

1 INTRODUCTION

The past decade has experienced rapid development in digital pathology, which transforms physical glass slides into high-resolution digital whole slide images (WSIs) [65]. This transformation lays the foundation for assisting diagnoses with machine intelligence [2, 15, 36], and might improve patient management ultimately [11]. To date, AI (Artificial Intelligence) has been proposed for a broad spectrum of potential applications of pathology [41, 67, 75, 83, 85], with some achieving performance on par with human beings in labs [10, 91]. Furthermore, various AI models have been adopted into tools to support pathologists' tasks, targeting automating parts of pathologists' workflow to reduce their examination burdens [17, 26, 53]. However, it is still challenging to convince pathologists to transform from manual diagnosis to AI-based methods in practice. We believe this is caused by the dichotomy between AI and medical communities — while the existing medical AI research focuses on improving performance, there is a lack of understanding of how doctors could benefit from AI and effectively use it for diagnosis [48, 60, 77, 90].

This onerous issue — the need to integrate AI-based tools into the medical workflow — has recently gained extensive attention in the HCI community. Empirical studies have interviewed medical professionals about their attitude toward using AI in practice, and suggest that medical systems should "state explicitly on how AI benefits users" [17] and "connect to existing clinical processes" [43]; it also indicates "unique difficulties" in converting human-AI interaction guidelines to tool support [90]. To this end, previous literature has explored the designs and influence of human-AI collaborative workflows for medical professionals [9, 30, 52, 84]. For pathology, numerous works have revealed the potential of human-AI collaborative systems to support doctors' exploration of one or more pathological patterns [16, 24, 53]. Extending the success of previous works, this work focuses on pathologists' more complicated diagnosis tasks, and studies how interfaces should be appropriately designed between pathologists and AI to address the workflow integration challenge, given the AI's incompatibility with existing pathologists' diagnosis workflow.

To reveal how AI-aided systems should be designed, we first conducted a formative study with four experienced pathologists (average experience $\mu = 21.25$ years) and summarized the main findings into the following design challenges:

(1) **Comprehensiveness**. Previous pathology decision support systems assist perspectives of pathologists' tasks, such as searching for one/more pathological patterns [53], or assisting adjudications on areas of interest [16, 38]. However, it is still challenging for the current systems to support diagnoses with multiple criteria from multiple pathological tests. This requires AI-aided pathology systems to comprehensively incorporate multiple criteria through a tight collaboration with pathologists;
(2) **Explainability**. Previous eXplainable AI (XAI) research interprets AI predictions using explainable elements, such as attention maps [91], concept attributions [16], and confidence scores [28]. However, it is still unclear how to effectively employ these components in pathologists' diagnosis, a time-sensitive but high-stakes process. In practice, pathologists expect to trace an AI-generated diagnosis to abundant evidence that explains such a decision;
(3) **Integrability**. Because of the complexity and the uncertainty of AI's output [89], it is challenging to present AI's comprehensive findings with explanations to match the diagnosis workflow of pathologists without incurring extra cognitive burdens, given the importance difference in each finding to the diagnosis according to the medical guidelines [56].

Building upon the design challenges from the formative study, we propose xPath — a comprehensive and explainable human-AI collaborative diagnosis tool that can assist pathologists' examinations integrated into their practice. Specifically, xPath can enhance pathologists' workflow integration with AI-based diagnosis from three aspects: *(i)* it reports multiple AI-computed pathology criteria, which are critical for diagnosis according to medical guidelines; *(ii)* it presents traceable evidence for each AI report, making it accountable and explainable;

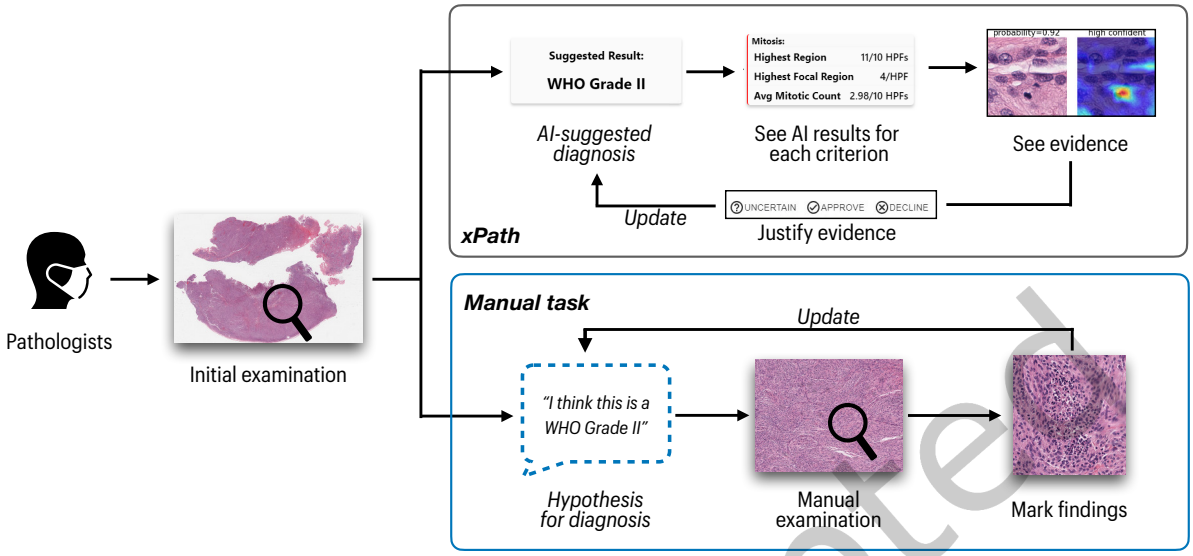


Fig. 1. Workflow of xPath (up): pathologists first see the AI-suggested diagnosis, then examine its results and evidence accordingly in an explainable manner, and examine the evidence to update the suggested diagnosis. This workflow follows a similar manual examination process of pathologists (down), which can improve AI's integration into pathologists' routine diagnoses.

(iii) it allows pathologists to perform diagnoses in a similar workflow to their routine practice (as shown in Figure 1).

We realize xPath with two design ingredients: **joint-analyses of multiple criteria** and **explanation by hierarchically traceable evidence**. First, the joint-analyses of multiple criteria present AI's findings based on multiple juxtaposed criteria from two pathology tests (Figure 2b), which are combined to produce a suggested diagnosis (Figure 2a) based on rules derived from the existing medical guideline [56]. Such a design addresses the comprehensiveness challenge, where pathologists are supported by AI-results of multiple criteria. Second, the design of hierarchically traceable evidence establishes a chain of accountable evidence for the diagnosis, explaining multiple levels of AI results, from high-level suggested diagnosis, to mid-level AI's reporting on each pathological pattern, and further to each piece of evidence: a user can trace the suggested diagnosis (Figure 2a) with a quantified score for the criterion (Figure 2d), to a list of evidence that contributes to the quantified score (Figure 2e), and further to examine each evidence with contextual information by registering it to the whole slide image (Figure 2f). Such a design addresses the explainability challenge by making the provenance of a criterion traceable and transparent. With the two designs, pathologists are freed from examining the pathology data with manual exploration of the high-resolution whole slide image, but building upon their diagnosis based on their seeing, understanding, and verifying AI results. Such a workflow with AI is also similar (and thus can be integrable) to pathologists' in practice (see Figure 1).

As for the validation of xPath, we hosted work sessions with twelve medical professionals in pathology[1] across three medical centers in the United States. We used data from a local medical center and asked our participants to diagnose with the same examination protocol as they had done in practice. We used working systems of

[1], which includes two attendings, two fellows, seven senior residents, and one junior resident.

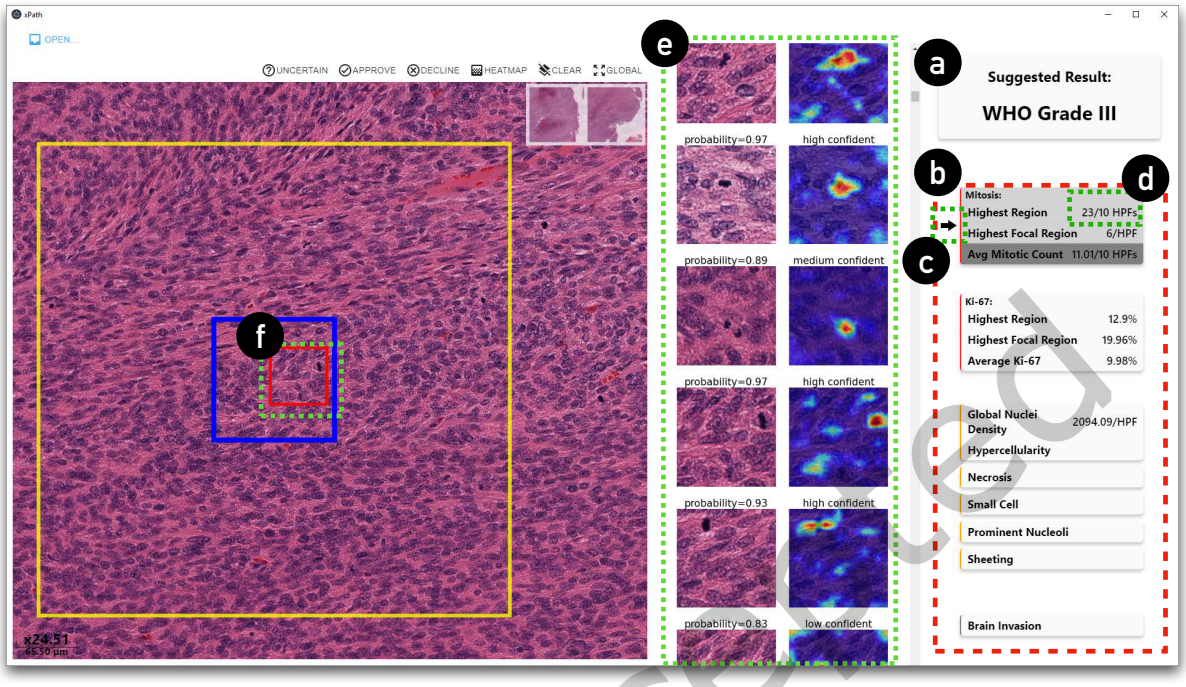


Fig. 2. xPath's interface design, illustrating the (a) suggested pathology diagnosis (*i.e.*, WHO Grade 3) with two key design ingredients of (b) joint-analyses of multiple criteria, where xPath offers comprehensive AI analysis of multiple critical pathology criteria for a diagnosis; explanation by hierarchically traceable evidence, explaining high-level suggested diagnosis to low-level AI-reporting on each pathological feature, including (c) an arrow that points to the deterministic criterion for the suggested diagnosis, (d) a quantified score for the criterion, (e) a list of evidence that contributes the quantified score, and (f) each piece of evidence registered to the whole slide image to support pathologists' examination with contextual information.

xPath and an off-the-shelf whole slide image viewer as the baseline. Our observations found that, with less than one hour's learning, participants could effectively utilize xPath to perform diagnosis. Specifically, they could use xPath's multi-criteria analysis by prioritizing one criterion and referring to others on demand. Furthermore, xPath's design of hierarchical explainable evidence enables participants to navigate between high-level AI results and low-level pathological details. A post-study questionnaire shows that, compared to the baseline system, participants reported xPath more integrable with their existing workflow (p=0.006, Wilcoxon rank-sum test, same below): they were more likely to use xPath in the future (p=0.002), and gave more overall preference on xPath (*i.e.*, 9/12 participants "totally prefer" using xPath than the baseline interface, and 3/12 "much more prefer" using xPath).

Benefiting from xPath's better workflow integration, participants reported xPath required less effort (p=0.002), and was more effective in reducing the workload (p=0.002) in performing diagnosis. Meanwhile, participants could make more accurate diagnosis decisions with xPath, where they gave 17/20 cases correct diagnosis using xPath, compared to 7/12 correct with the baseline interface.

1.1 Contributions

Our main contribution is two-fold: *(i)* throughout interviews with experienced pathologists, we identified their challenges in practice, and summarized that comprehensiveness, explainability, and integrability are the three key components for incorporating AI models into pathologists' workflow; *(ii)* based on the empirical findings, we proposed a human-AI diagnosis tool — xPath — that facilitates pathologists' routine examinations collaboratively, validated by a study that evaluates pathology professionals' diagnoses compared with a baseline system. Our study and findings shed light on how HCI researchers can design integrable AI-assisted systems to bring advancements to doctors' workflow.

2 RELATED WORK

In this section, we review the related work of xPath from three areas: *(i)* AI algorithms for processing pathology images, *(ii)* enhancing AI's workflow integration for medical applications, and *(iii)* human-AI collaborative tools for pathologists.

2.1 Processing Pathology Images with Data-Driven AI

With the recent development of digital pathology techniques, a considerable amount of datasets have grown around the theme of marking pathological patterns from digital pathology slides. Current datasets are primarily based on H&E (*i.e.*, Hematoxylin and Eosin, a type of pathology staining) slides, the most commonly used stained slides for providing a detailed view of the tissue. To date, these datasets cover a broad range of pathology practices, from conducting high-level diagnostic tasks, such as identifying breast cancer metastasis [54], to seeing low-level pathological patterns, such as mitoses [68, 79].

Such an increase in data availability in digital pathology has triggered a recent surge of data-driven techniques in a broad range of applications, such as screening negative biopsies [26], carcinoma detection [3, 8, 10], quantification of pathological features [23, 34, 80], and tumor grading [5, 27]. It is noteworthy that some previous AI models have achieved performance on par with human beings in lab studies. For example, Zhang *et al.* combined multiple neural networks, including a Convolution Neural Network (CNN), a fully connected neural network, and a Recurrent Neural Network (RNN), to diagnose urothelial carcinoma, which achieves matching diagnosis performance compared to a group of pathologists [91].

Besides H&E slides, AI algorithms have also been devised for other pathology tests that can assist decision-making, *e.g.*, Ki-67 immunohistochemistry (IHC) tests. For example, Xing *et al.* trained a fully connected convolutional network to perform nucleus detection and classification from Ki-67 slides [87]. In more recent research, Ghahremani *et al.* trained a cycleGAN network with more-precise immunofluorescence data as ground truth to improve cell-level semantic segmentation for IHC tests [32].

Although its broad applications and promising performance, data-driven AI in pathology has caused rising ethical concerns because of the high-stakes nature of performing diagnoses [20]. Multiple works ask AI to provide algorithm transparency [17] and result accountability [61, 72]. And several studies have included eXplainable AI (XAI) techniques to improve the transparency of data-driven AI for pathology. For example, Gehrung *et al.* employed the saliency map visualization to highlight the spacial support for the model prediction, and suggest that the saliency maps a strong agreement with pathologists' labels [31]. To investigate pathologists' attitudes towards XAI elements, Evans *et al.* further conducted a user-oriented study and found that simple visual explanations were preferred because they were closer to pathologists' visual examinations on the slides [28].

Going beyond providing explanations for AI predictions, other works aim to build interpretable AI for healthcare applications. For example, Choi *et al.* mimicked physicians' practice of examining electronic health records and introduced an interpretable RNN model that diagnosed by detecting patients' past visits [22]. Koh *et al.* trained a concept bottleneck model that can classify X-ray images with human-understandable concept values as

interpretations [49]. However, it is noteworthy that interpretable AI in the pathology imaging domain is not as popular as in general AI research. We believe this is partly related to AI training: first, interpretable AI usually requires human-annotated labels (*e.g.*, concepts) for training, while pathologists' annotations are hard to acquire [71]; second, it adds difficulties to training interpretable AI because its additional interpretable constraints [70].

The progress of the AI and XAI techniques has built fundamentals of using AI to automate pathologists' tasks without losing transparency. However, the main focus of AI in the medical domain is to improve performance, while XAI research targets to explain AI findings. We argue that it is insufficient to assist pathologists' diagnoses by only optimizing the AI algorithms or simply applying XAI designs. This is because their poor integration into the medical workflow might add burdens to pathologists, which disincentivizes them to use AI systems in practice [90]. In this work, we seek a better understanding of pathologists' expectations of AI by working closely with a group of pathologists. Based on which we further conclude three design requirements for pathology AI systems — comprehensiveness, explainability, and integrability — to enhance the integration of AI-aided systems in pathology.

2.2 Enhancing AI's Workflow Integration for Medical Applications

In the history of medical AI systems, workflow integration has been recognized as a key value for medical users. For example, Teach *et al.* have studied physicians' attitudes toward clinical consultation systems and offered suggestions on computer-based decision support systems, *e.g.*, "minimizing changes to current clinical practices" [76]. Middleton *et al.* have reviewed research on clinical decision support systems since 1990 and pointed out that the poor integration in clinicians' workflow is becoming a barrier preventing the application of such tools [62]. Yang *et al.* indicated that a medical AI tool should set the explicit goal of helping medical users increase the overall quality of examination, instead of insufficiently automating a part of their work [90].

In the general healthcare domain, literature has attempted to enhance workflow integration by improving medical users' engagement in the design process of AI systems. For example, Sendak *et al.* included medical professionals in designing and implementing a deep-learning-driven sepsis monitoring system. Based on the co-designing process, they summarized takeaways to improve workflow integration, including "respect professional discretion" and "create ongoing feedback loops with stakeholders" [72]. Jacobs *et al.* further concluded that medical systems should "offer on-demand explanations" to address the mismatch between AI predictions and the medical guidelines [43].

Numerous studies have explored the potential usage, issues, and influence of employing human-AI collaborative workflows in clinical settings. For example, Beede *et al.* studied socio-environmental factors that influenced AI performance, nurses' workflow, and patient experience while using a deep learning system for diabetic eye disease [9]. Wang *et al.* revealed challenges of usability, technical limitations, and human trust that emerged from applying an AI-powered clinical diagnostic support system [84]. Fogliato *et al.* discovered that demonstrating AI inference at the start of radiologists' reading of X-ray images would increase doctors' agreement [30]. Lee *et al.* reported that the human-AI collaborative system could increase therapists' agreement on the rehabilitation assessment [52].

Narrowing down to the pathology domain, Cai *et al.* highlighted pathologists' needs for information from AI, which included the AI's capabilities measured in well-defined metrics and transparency to overcome subjectivity [17]. Gu *et al.* summarized six design lessons for interactive AI systems in pathology, suggesting AI systems in pathology should "provide the actionability of the AI guidance" and "narrow down to small regions of a large task space" [35].

The design conclusions and guidelines open up opportunities to enhance AI's integration into pathologists' workflows. However, there are still limited working tools that support pathologists' diagnoses in the wild, consisting of examining multiple criteria from multiple pathology tests. In this work, we aim to enhance AI's

workflow integration using the task of meningioma (a type of brain tumor) grading as a case study. Specifically, we propose two designs for pathology AI systems: **joint-analyses of multiple criteria** and **explanation by hierarchically traceable evidence**. We observe how pathologists interact with these two designs and summarize recurring themes, providing first-hand information for future pathology AI system designs.

2.3 Human-AI Collaborative Tools for Pathologists

One way to increase AI's workflow integration for medical users is by enabling them to collaborate with AI. And enabling human-AI collaboration requires "goal understanding, preemptive task co-management and shared progress tracking" [82].

Recent HCI research has demonstrated numerous examples of human-AI collaboration in various general tasks, such as content creation [44, 45, 86], design [21, 51], well-being [88], and accessibility [55]. For medical tasks, various human-AI collaboration systems have shown their validity in improving doctors' agreement [15, 30, 52], mental effort [16], and accuracy [12]. However, literature has also suggested that AI performance might be influenced by clinical factors in the wild [9, 66, 84]. In the pathology domain, multiple works have shown that the human + AI approach could potentially increase the quality of diagnoses. For example, Wang *et al.* reported that combining AI and human diagnoses improves pathologists' performance in breast cancer metastasis classification with an ~85% reduction in human error rate [83]. More recent work by Bulten *et al.* has suggested that the introduction of AI assistance increases pathologists' agreement with the expert reference standard in prostate cancer grading [15].

A number of existing human-AI collaboration projects on pathology have been focused on Content-Based Image Retrieval (CBIR). With a given slide (or patch) from pathologists, such tools retrieve image examples of a similar pattern to help the decision-making. For instance, Hegde *et al.* proposed a reversed image searching tool to help pathologists find image patches with similar pathological features or disease states [38]; Cai *et al.* enabled pathologists to specify custom concepts that guide the retrieval of similar annotated patches of pathological patterns [16]. However, the CBIR focuses on image searching: what images to search, how to use the search results, and what to conclude according to searching results. On the other hand, diagnosing/grading carcinoma in digital pathology is more complicated, requiring pathologists to detect multiple pathological features and aggregate them according to medical standards for decision-making. And xPath is considered a tool of Computer-Aided Diagnosis (CAD) or Clinical Decision Support System (CDSS).

Existing CAD/CDSS tools can enhance the detection in digital pathology with visualization. For example, Corvo *et al.* developed PathoVA, which provided AI support for breast cancer grading by visualizing three types of clues [24]. The system could also track pathologists' interactions and help them generate reports. Krueger *et al.* enhanced users' exploration of multi-channel fluorescence images to support cell phenotype analysis [50]. Specifically, the tool maintained hierarchical statistics about the number of cell-level findings to help a user keep track of analysis and interactively update the statistics with machine learning algorithms on the fly. These tools provide a bottom-up approach to assist pathologists in making a diagnosis: pathologists are only prompted with low-level AI-generated clues (*e.g.*, highlighting tumor cells with a segmentation map); then, the diagnosis is drawn by pathologists from fusing observations with these clues. In contrast, xPath allows pathologists to evaluate a case with a top-down approach: they can first see an overall grading (top-level) based on joint analyses of multiple criteria, then drill down to localized areas with traceable evidence and further to low-level patterns for verification and correction. Such a design is similar to pathologists' examining the image manually, where they first develop hypotheses and interactively refine them by adding supporting evidence.

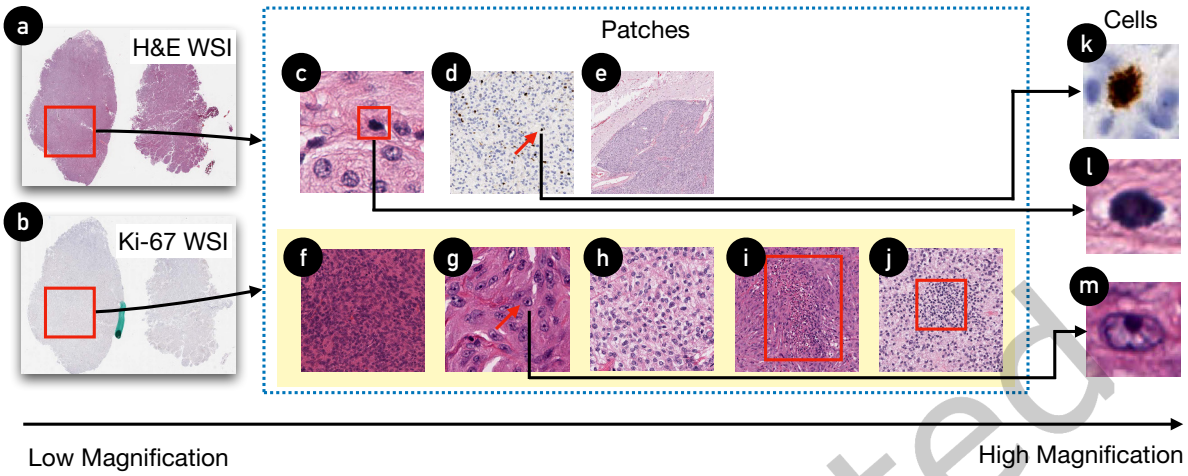


Fig. 3. Examples of criteria used for the meningioma grading. (a) The resected tissues are first stained with H&E solution. (b) An additional Ki-67 IHC test is usually used to locate mitoses. According to the WHO grading guidelines, pathologists look for (c) mitotic cells (marked in the red box) in high-power fields with the help of (d) Ki-67 stains; (e) brain invasion (invasive tumor cells in brain tissue); five pathological patterns, including (f) hypercellularity (an abnormal excess of cells), (g) prominent nucleoli (enlarged nucleoli pointed by the arrow), (h) sheeting (loss of 'whirling' architecture), (i) necrosis (irreversible injury to cells marked in the red box), (j) small cells (tumor cell aggregation with high nuclear/cytoplasmic ratio marked in the red box). For some criteria, *e.g.*, mitosis (k,l) and prominent nucleoli (m), pathologists are required to zoom further into the high magnification level for examination.

3 MEDICAL BACKGROUND

In this work, we target the task of meningioma (a type of brain tumor) grading as a case study to probe the design of human-AI collaborative tools for pathology diagnosis. The meningioma grading is selected because of its complexity — it covers three aspects of difficulties for pathologists: *(i)* multiple morphological and immunohistological features utilizing at least two kinds of pathology tests (*i.e.*, Hematoxylin and Eosin (H&E) slides and Ki-67 immunohistochemistry (IHC) tests) for the grading of the tumor, *(ii)* alternate high and low magnification images to detect large structures (*i.e.*, brain invasion, see Figure 3e) or small events (*i.e.*, mitosis, see Figure 3c), and, *(iii)* examine the entire tumor (occasionally as many as 20 or more slides) for frequently rare features (*i.e.*, spontaneous necrosis, see Figure 3i). As such, the practice of grading meningiomas is a favorable arena for studying how human-AI collaborative systems should be designed to assist pathologists in carrying out multiplex tasks.

According to the World Health Organization (WHO) guidelines (2016), meningiomas can be graded as Grade 1, Grade 2, or Grade 3 [56]. The current grading of meningioma in the new WHO guideline (2021) still recommends the same criteria for grading, although the nomenclature is slightly different. Additionally, new molecular alterations are added to determine the tumor grade [57].

The accurate grading of meningioma is vital for treatment planning: the Grade 1 tumors can be treated with either surgery or external beam radiation, while Grade 2/3 ones often need both treatments [81]; meanwhile, research shows that patients with Grade 3 meningiomas suffer a higher recurrence rate as well as lower survival rate in comparison to Grade 2 patients [64].

ID	Occupation	Years of Experience	Familiarity of Meningiomas
FP1	Attending/Professor	44	Examine Weekly
FP2	Attending/Assistant Professor	22	Examine Weekly
FP3	Attending	10	Examine Weekly
FP4	Attending	9	Examine Weekly

Table 1. Demographic information of the participants in the formative study.

Pathologists need to search and locate multiple pathological features across various magnifications with optical microscopes or digital interfaces in order to determine the tumor grade. Specifically, they first localize the regions of interest (ROIs) in low magnification (x40), then switch to the patch level with a higher magnification (x100), and sometimes zoom further with the highest magnification (x400) to examine cellular architecture. These steps are usually repeated multiple times until pathologists have collected sufficient findings to conclude a grading and sign out the case.

Figure 3 briefly visualizes examples of pathological features that pathologists need to find. Pathologists' work starts with the H&E slides (Figure 3a). Apart from the H&E, Ki-67 IHC tests [1] are often used (Figure 3b) to provide an estimated proliferation index (Figure 3d,k), which is highly correlated to meningioma grading. According to the WHO guidelines[2] [56, 57], grading meningiomas is based on the findings of multiple microscopic or large-sized pathological features. As such, meningioma grading is challenging and high-stakes — an overestimated study would incur unnecessary treatment on patients, and an overlooked one would cause a delay of necessary treatment.

4 FORMATIVE STUDY

We conducted a formative study to reveal the system requirements for human-AI pathology diagnosis. Specifically, we recruited four experienced pathologists (average experience $\mu = 21.25$ years) from a local medical center through word-of-mouth. All participants had examined meningiomas weekly. The demographic information of the participants is shown in Table 1. Two out of four participants (FP3, FP4) have used digital pathology systems, and the primary software they used is Imagescope[3]. For familiarity with AI, one participant knows machine learning, one has passing knowledge, and two have little.

As for the process of the formative study, we started by describing the project's motivation and presented participants with a real meningioma whole slide image. Next, we asked the participants to examine the case and encouraged them to talk aloud about their examination process. We followed up with a semi-structured interview and let the participants describe the challenges in their practice and their expectations of an AI-enabled system to assist such a process. The average duration of the semi-structured interviews was about 25 minutes, and the average length of the study was about 60 minutes. Please refer to the supplementary material for the moderator's guide in the semi-structured interview.

4.1 Existing Challenges for Pathologists

We first transcribed the audio recordings of all interviews. One experimenter coded the transcripts and shared the recurring challenges mentioned by the participants. A second experimenter coded individually and took a pass on the first experimenter's findings. Then, a third experimenter joined to discuss with the previous

[2]Please refer to the Appendix A for detailed descriptions of the WHO guidelines for meningioma grading.
[3]https://www.leicabiosystems.com/us/digital-pathology/manage/aperio-imagescope/

two experimenters and resolved the disagreements. Resulting from the complicated the medical guideline, we discovered three challenges in the current pathology practice of meningioma grading:

Time Consumption. The small-scaled characteristics in the patterns of interest and the very high resolution of slides make the meningioma grading highly time-consuming for pathologists. A resected section from a patient's brain tissue would generate eight to twelve H&E slides, and pathologists need to look through all those slides and integrate the information found on each slide. Except for the few experienced pathologists, meningioma grading can be time-consuming to go through because a single patient's case often consists of 10+ slides — "*If you don't see obvious features of malignancy, like necrosis or mitosis, you have to search all of the slides in high power to look for mitosis, which will take a few hours*" (FP4) Automating portions of the slide examination process by AI can potentially reduce such time consumption, alleviate pathologists' workload, and increase the overall throughput.

Subjectivity. There are high intra- and inter-observer variations during the grading of tumors. Pathologists summarize three factors contributing to such subjectivity: *(i)* a lack of precise definitions — the WHO guidelines do not always provide a quantified description for the five pathological features of high-grade meningioma. For example, for the 'prominent nucleoli' criterion, the WHO guideline does not specify how large the nucleolus should be considered as 'prominent', described by FP2 — "*... small cells, large nucleoli ... nobody has defined what that means...*"; *(ii)* implementation of the examination process — for example, the mitotic count for grade 2 meningioma is defined as 4 to 19 mitotic cells in 10 consecutive high-power fields (HPFs)[4]. However, the guideline does not specify the sampling rules of these 10 HPFs. As a result, different pathologists are likely to sample different areas on the slide; *(iii)* natural variability in people, such as the level of experience, time constraint, and fatigue [25] — "*One person would like to say it is mitosis, while the other person would say 'not really', because it is not good enough.*"(FP4) For AI, the definition and implementation of guidelines can be codified into the model and visualized in the system that performs consistently to overcome people's variability.

Multi-Tasking. Going beyond the time consumption and subjectivity, participants also mentioned that it was also challenging for less-experienced pathologists to "multitask", *i.e.*, cross-referencing amongst multiple criteria at the same time, rather than going through one after another sequentially. The "multitasking" operation is challenging because it requires pathologists to memorize which criterion they had found and where they were simultaneously. However, we believe such a limitation can be addressed by introducing digital systems without AI, where computers can memorize pathologists' previous annotations and interactions.

4.2 System Requirements for xPath

Regarding pathologists' expectations about the system, we summarized three requirements to enhance workflow integration: comprehensiveness, explainability, and integrability. Note that participants also expect the AI to be accurate and reproducible for meningioma grading — "*If the machines cannot provide accurate material, it is not a worthwhile system ... It would be good if two different machines can give the similar quality of mitosis.*" (FP1) However, instead of including them in the *system* requirements, we believe such concerns can be addressed by the introduction of high-performance AI, which we will demonstrate in Section 6.

Comprehensiveness. According to the current medical guideline, the grading of meningiomas involves multiple sources of pathology tests (from H&E and Ki-67) and criteria (*e.g.*, mitosis, necrosis, brain invasion). To incorporate xPath into the current practice, the system should comprehensively, systematically, and exhaustively support all these pathology tests and criteria to ensure that pathologists do not miss crucial findings.

Explainability. In lieu of a single grading result from a black-box AI model, the system should provide visual evidence to justify the AI's findings according to the medical definition of the criterion. This is because some criteria (only visible under high magnifications) requires examining lower-level details in order to interpret an AI's finding and further needs to be traceable to the original location in the whole slide image for a review with

[4]The size of field-of-view under x400 magnification of a optical microscope.

more contextualized information. Overall, there should be explainability both globally (how results from multiple criteria are combined to yield a grading) and locally (which includes *(i)* what evidence leads to the computed result of each criterion, *e.g.*, where mitoses are detected that lead to the number of mitosis counts, and *(ii)* why a specific piece of evidence is captured by AI, *e.g.*, which part of the evidence convinces the AI that it contains mitoses).

Integrability The system should allow pathologists to diagnose with AI similar to their daily routines of manual examination. Specifically, the system should first suggest a hypothesis for diagnosis and provide evidence to support it. Meanwhile, given that errors are inevitable for most existing AI models, the system should allow pathologists to refine AI's findings by retrieving detailed contextualized evidence on demand. When showing the evidence of grading, the system should not overwhelm pathologists with all evidence from a whole slide; rather, it should direct pathologists to the representative regions of interest. Finally, the system should enable pathologists to cross-check each criterion and override the results manually when they detect an error.

5 DESIGN OF XPATH

Guided by the aforementioned system requirements, we developed xPath with two key designs for pathology AI systems: *(i)* joint-analyses of multiple criteria and *(ii)* explanation by hierarchically traceable evidence. We first detail the two designs and then describe how a pathologist uses xPath to perform a meningioma grading task.

5.1 Joint-Analyses of Multiple Criteria

Based on the formative study, we found that pathologists rely on the WHO meningioma grading guideline for meningioma grading [56] involving multiple criteria. Thus xPath's design follows the WHO guideline and employs AI to compute eight critical criteria for meningioma grading[5]. Details on the AI implementation are described in Section 6. These criteria can be split into two categories: quantitative and qualitative. For the quantitative criteria (*i.e.*, mitotic count, Ki-67 proliferation index), we show their predicted *quantitative values* directly. For the other criteria dealing with the *presence* or *absence* of a specific pathological pattern, xPath provides recommendations of regions of interest (ROI) hotspots according to the largest aggregations of AIs' probabilities.

Figure 4 demonstrates the interface of multiple criteria, which shows the current suggested grading for the tumor (*i.e.*, the suggested 'WHO grade 2', Figure 4a) and a structured overview of each criterion (Figure 4b). xPath displays an arrow to indicate the main contributing criterion (Figure 4c), the most deterministic AI findings for the suggested diagnosis, according to the meningioma grading guidelines (see Appendix A). For example, in Figure 4, xPath suggests the "*mitotic count*" is the main contributing criterion, because it has detected 12 mitoses in 10 high-power fields (HPFs) (Figure 4c, highest region). Such AI findings directly satisfy descriptions of WHO grade 2 meningiomas, making "*mitotic count*" the main contributing criterion. Going beyond the main contributing criterion, all the criteria are linked with the evidence or regions of interest related to the findings. Moreover, AI's recommendation on all the criteria can be overridden by the pathologist (Figure 4d). And xPath uses color bars (Figure 4e,f,g,h) to indicate the status.

In summary, the joint-analyses of multiple criteria addresses the challenge of comprehensiveness by providing important information for pathologists according to the medical guideline. xPath also achieves global explainability by presenting how different AI-computed criteria are combined to arrive at a diagnosis. Such a design can enhance AI's workflow integration because it exposes the pathologist to high-level AI findings when

[5]... which includes the mitotic count, Ki-67 proliferation index, hypercellularity, necrosis, small cell, prominent nucleoli, sheeting, and brain invasion. Note that this work does not consider using AI to identify the subtypes (*e.g.*, clear cell, frank anaplasia) because we believe they are relatively easier to be discovered and judged by pathologists.

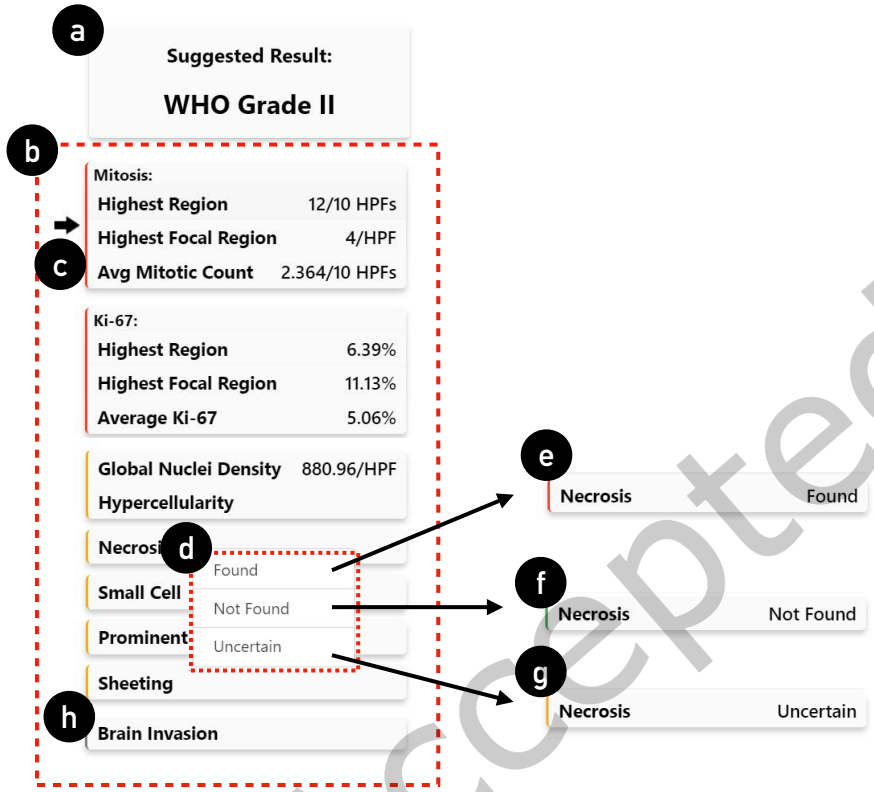


Fig. 4. Joint-analyses of multiple criteria in xPath's design: (a) the overall suggested grading; (b) a structured overview of each WHO criterion with (c) an arrow highlighting the main contributing criterion to the suggested grading; (d) users can override criteria by right-clicking on each item and change the result to 'found', 'not found' or 'uncertain'; xPath provides color bars to indicate the status of each criterion: (e) red indicates a confirmed abnormal criterion (or *presence*), (f) green indicates a confirmed normal criterion (or *absence*), (g) orange indicates the criterion is unconfirmed/confirmed uncertain, and (h) gray indicates the criterion is not applicable in this case.

they onboard the case. As such, they can establish an initial understanding and develop hypotheses, which also facilitates them to double-check with their examination later.

5.2 Explanation by Hierarchically Traceable Evidence for Each Criterion

Another finding from the formative study is that, besides a global explanation of the overall grading, pathologists also would like to see evidence that justifies AI's grading, *e.g.*, how AI processes the image of a local patch (for local explainability). Hence, we designed xPath to provide such explanations by hierarchically traceable evidence: xPath enables pathologist users to examine and justify the evidence with a top-down human-AI collaboration workflow. Specifically, at the **top level**, pathologists can first see the suggested diagnosis recommended by xPath (Figure 5a). Then, they can continue to dive down and examine a list of AI-computed criteria (Figure 5b). Each criterion can be boiled down to a list of **mid-level** samples (Figure 5c). For the most important criterion — mitosis, xPath demonstrates a series of explanations in each sample, including AI's output probability (Figure

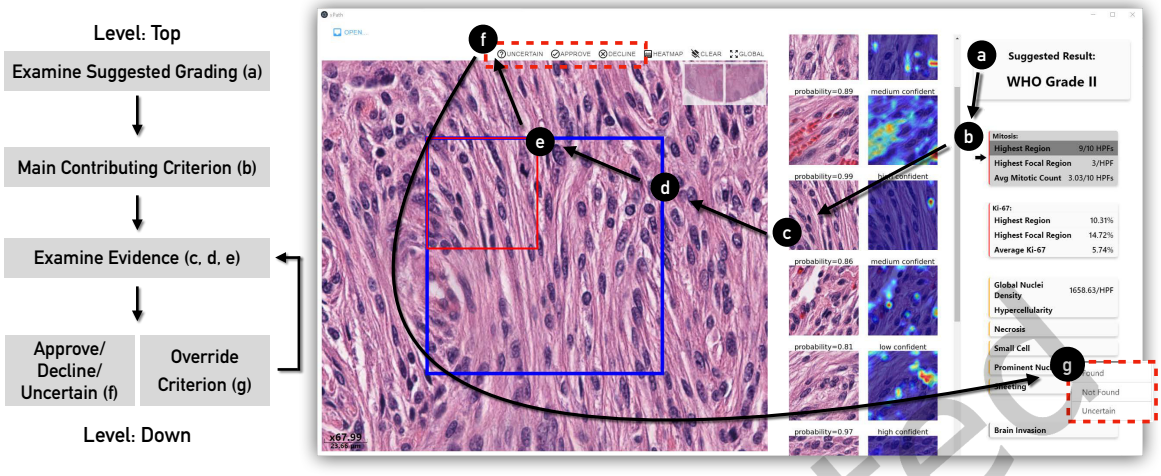


Fig. 5. xPath presents a top-down human-AI collaboration workflow for pathologists to interact with xPath (left) and pathologists' corresponding footprints on the xPath's frontend user interface with examining the mitosis criterion as an example (right). A pathologist user starts from (a) the AI-suggested grading result and then examines (b) the main contributing criterion. They can further examine (c) the evidence list, and register back into the original whole slide image in higher magnifications (d,e). Furthermore, users can (f) approve/decline/declare-uncertain on the evidence, or (g) override AI results directly by right-clicking on each criterion. Users might repeat the same workflow (c-g) multiple times to examine other criteria (one criterion for each time). Meanwhile, xPath's suggested grading (a) will be updated as the user justifies AI's findings. The user may continue to interact with xPath until they have collected sufficient confidence for a diagnosis.

6a), AI's confidence level (Figure 6b), and a saliency map (Figure 6c) that highlights the spatial support for the mitosis class in the reference image[6], allowing pathologists to check AI's validity on each sample quickly. Further, at the **low-level**, xPath supports registering each sample into the whole slide image (WSI) to enable pathologists to examine with higher magnification and search nearby for more contextual information (Figure 5d,e).

With the provided **mid-** and **low-level** information, a pathologist can approve/decline/declare-uncertain a sample for a criterion with one click (Figure 5f), or directly override AI's results on each criterion (Figure 5g). Correspondingly, the overall suggested grading (Figure 5a) is updated dynamically upon the user's input. Such a diagnosis-contesting workflow allows pathologists to challenge AI's suggested diagnosis by seeing AI's reasoning line and evidence, which increases the "contestability" as described in previous HCI research in healthcare [39].

Such a workflow mimics a scenario that we found in the formative study: pathologists might assign low-level tasks (*e.g.*, marking ROIs, finding specific criteria) to trainees in practice. They can continue to perform a differential diagnosis (*i.e.*, building hypotheses and ruling out less-probable cases with findings) based on trainees' reports. By replacing trainees with AI, we emulated the relationship between the pathologists and trainees, thus making AI integral to pathologists' current practices.

Figure 7 demonstrates typical examples of evidence provided by xPath. Particularly, for the mitosis-related criteria (*i.e.*, mitotic count from H&E WSI and Ki-67 proliferation index from Ki-67 IHC WSI), which are commonly used for meningioma grading, we introduce two 'shortcuts' for pathologists to look into AI's results:

- **Highest Region Sampling**. One WHO criterion is the mitotic count in 10 consecutive high-power fields (HPFs). Our formative study found that the inter-observer consistency of "10 consecutive HPFs" is low due

[6]Please refer to the supplementary material for the implementation of calculating the confidence level and the saliency map.

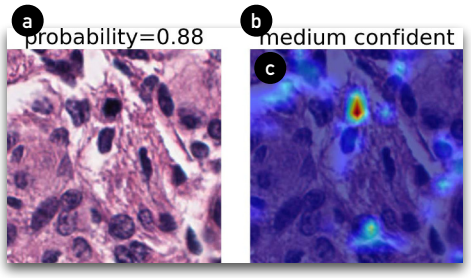


Fig. 6. For the mitosis criterion, xPath demonstrates a series of explanations in each mid-level sample, including the (a) AI's probability, (b) AI's confidence level, which is calculated by the probability thresholds, and (c) a saliency map (calculated by the Grad-CAM++ algorithm [19]) that highlights the spatial support for the mitosis class in the reference image on the left.

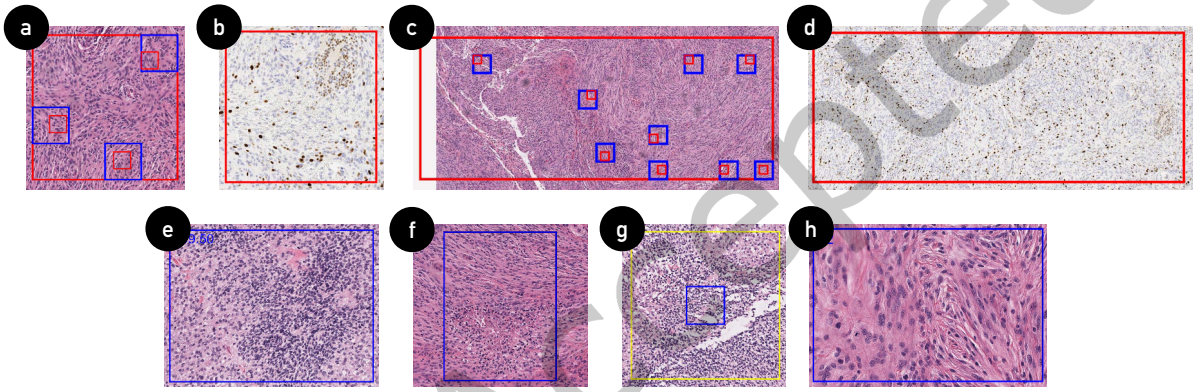


Fig. 7. Selected pieces of sampled evidence: (a) a highest focal region sampling result of mitotic count on H&E slide (red box, 1HPF), the small blue frames indicate the rough positions of detected mitoses, and the smaller red boxes in the blue frames mark the positions of mitoses (that are shown on the evidence list) found by xPath's AI; (b) a highest focal region sampling result on the Ki-67 IHC slide (red box, 1HPF); (c) a highest region sampling result of mitotic count on H&E slide (red box, 10HPFs) with mitoses reported by xPath's AI (the blue frames and smaller red boxes); (d) a highest region sampling result on the Ki-67 IHC slide (red box, 10HPFs); (e) a hypercellularity ROI sample (blue box); (f) a necrosis ROI sample (blue box); (g) a small cell ROI sample (the inner blue box, the outer yellow box marks the dimension of 1HPF); (h) a prominent nucleoli ROI sample (blue box).

to the difference in the ROI sampling rules adopted by pathologists. To address this problem, xPath provides the highest region sampling tool. The highest region is defined as a 2×5 HPF area with the highest number of mitotic counts (Figure 7c) or the highest Ki-67 proliferation index (Figure 7d). This tool speeds up a pathologist's work by helping them locate 10 consecutive HPFs as required by the WHO guidelines.

- **Highest Focal Region Sampling**. From our formative study, pathologists mentioned that high-grade meningiomas share a common feature of increased mitotic activities in a localized area. Hence, xPath provides the highest focal sampling tool to help pathologists better localize highly concentrated mitosis/Ki-67 proliferation index areas. In xPath, the highest focal region is calculated as the one HPF with the highest number of mitotic counts (Figure 7a) or the highest Ki-67 proliferation index (Figure 7b). Using this tool, pathologists can locate foci of highly-mitotic areas that the highest region sampling might miss.

Pathologists can go beyond the sampled areas and navigate the high-heat areas using heatmaps generated for the whole slide (please see the supplementary material for details). For example, the mitosis heatmap registers all AI-detected positive mitotic cells as a mitotic density atlas, where high-heat areas indicate a high density of mitotic cells. As such, the heatmap would serve as a 'screening tool' to help pathologists filter out unrelated areas and rapidly narrow down to the ROIs that are scattered in an entire WSI. xPath provides such 'screening tools' for all criteria.

After pathologists have finished examining one criterion, they can proceed to justify the rest of the criteria with the same top-down workflow (one iteration for each criterion). During such an iterative process, xPath will update AI's findings on an individual criterion and, if necessary, the overall suggested grading as well. Finally, pathologists can make a diagnosis once they have collected sufficient confidence for the grading diagnosis.

In summary, in contrast to prior work that enables pathologists to define their own criteria for finding similar examples [16], xPath aims at making examinations based on an existing criterion traceable and transparent with evidence, which allows pathologists to see and understand why AI derives such findings. Furthermore, pathologists can challenge (or "contest" [39]) these AI findings with a top-down workflow to refine the suggested grading diagnosis. Such collaboration between pathologists and AI is similar to that with pathology trainees, where pathologists can perform a differential diagnosis based on trainees' findings.

6 IMPLEMENTATION OF XPATH'S AI BACKEND

xPath implements an AI-aided pathology image processing backend to compute the eight pathological criteria of the mitotic count, Ki-67 proliferation index, hypercellularity, necrosis, small cell, prominent nucleoli, sheeting, and brain invasion. In this section, we briefly describe datasets, the AI processing pipeline, and AI training details. Finally, we report the performance of each of the AI models from a technical evaluation.

6.1 Processing WSIs with AI

xPath aims to screen the entire whole slide image (WSI) using AI and then determine suggested grades based on the AI findings. To achieve this, xPath includes six AI models and two rules, one for each criterion, to general initial AI results. For each WSI, we first used a sliding window technique to cut it into smaller tiles. For each tile, we further employed a series of AI models to calculate six criteria (*i.e.*, nuclei count (Figure 8f), necrosis probability (Figure 8g), sheeting probability (Figure 8h), mitosis (Figure 8i), prominent nucleoli (Figure 8j), and Ki-67 proliferation index (Figure 8m)). Based on the AI-computed nuclei count, we further used two rules to support the reporting of the small cell and the brain invasion patterns. xPath can recommend small cell tiles based on the nuclei count of each tile (Figure 8k). Furthermore, the brain invasion was visualized by classifying the brain *vs.* tumor regions according to the nuclei count (Figure 8l). This is because meningioma tumor areas usually have a high nuclei density, while normal brain tissues are not. After the AI models had processed each tile, xPath calculated the ROIs using a set of rules. Please refer to the supplementary material for more detailed descriptions of xPath's AI implementation and the ROI generation process.

6.2 Dataset and Model Training

Since there were no pre-trained models nor public meningioma datasets for the pathology patterns of mitosis, necrosis, prominent nucleoli, and sheeting, we built an in-house dataset consisting of 30 WSIs (WSI total size = ~ 54.9 GB) from a local medical center to train AI models to classify these four patterns. The WSIs were scanned by an Aperio CS2 scanner in x400 magnification (pixel size=0.25μm). The ground truth labels were collected in two ways: *(i)* for the mitosis, the pathologist labeled with an online labeling system; *(ii)* for other criteria, the pathologist marked ROIs using the Imagescope software. We then cropped the labeled ROIs with a random-crop technique, and the tiles in different sets were generated from a different group of ROIs. In sum, the final dataset

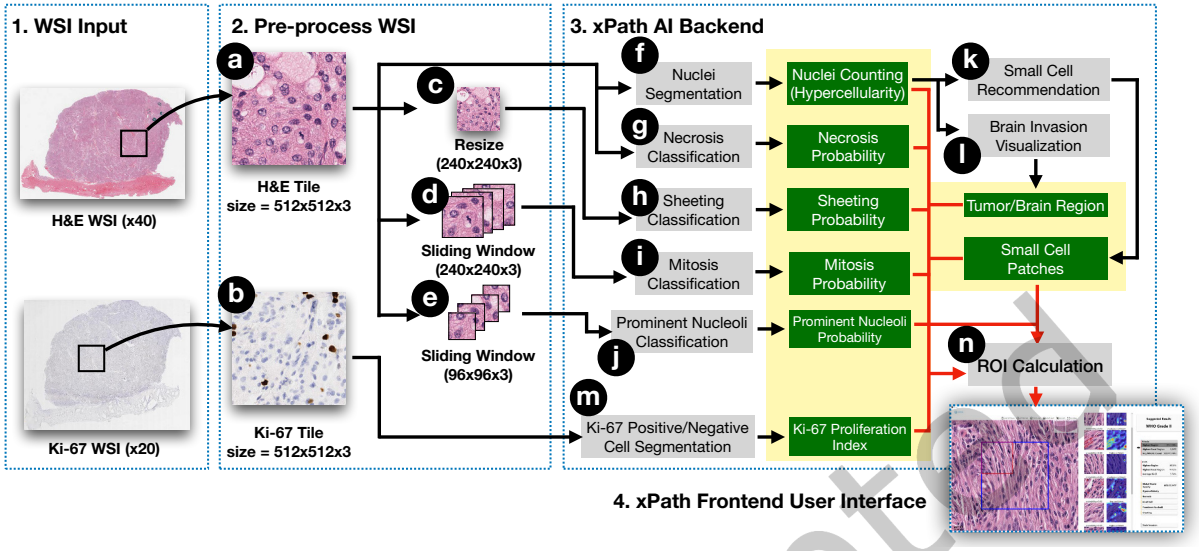


Fig. 8. Data processing pipeline of xPath: *(i)* xPath takes H&E and Ki-67 whole slide images (WSIs) as input. *(ii)* For each WSI, xPath uses a sliding window method to acquire (a) H&E and (b) Ki-67 tiles; Furthermore, each H&E tile is processed with (c) resizing, (d) sliding window ($240 \times 240 \times 3$), and (c) another sliding window ($96 \times 96 \times 3$) to fit the inputs of the down-stream AI models. *(iii)* xPath's AI backend takes over the pre-processed tiles and employs multiple AI models to detect WHO meningioma grading criteria from each tile. Given an H&E tile, xPath uses (f) a nuclei segmentation model to count the number of nuclei (for hypercellularity judgment), (g) a necrosis classification model to calculate necrosis probability, and (h) a sheeting classification model to calculate sheeting probability. xPath further utilizes the nuclei counting results for (k) small cell recommendation, and (l) brain invasion visualization. For a $240 \times 240 \times 3$ tile, xPath uses (i) a mitosis classification model to obtain the mitosis probability. For a $96 \times 96 \times 3$ tile, xPath uses (j) a prominent nucleoli classification model to predict prominent nuclei probability. For each Ki-67 tile, xPath (m) detects positive and negative nucleus to calculate the Ki-67 scores; *(iv)* xPath further (n) calculates ROIs based on all AI-computed results (marked in the green boxes), and shows them as evidence on the frontend user interface for pathologist users to justify.

has a size of ~ 16.1 GB. It consists of four training and testing sets, covering the four pathology patterns (as shown in Table 2).

To train the models, for each criterion, we further randomly selected a subset of the training set to be the validation set. Specific thresholds were decided by the maximum F1 scores achieved by each model in the validation set. Please find the supplementary material for more specific training details.

6.3 Technical Evaluation

We report the performance of AI models on testing sets. Specifically, we test the supervised models for recognizing mitosis, necrosis, prominent nucleoli, and sheeting criteria, and report the Precision-Recall curve, as shown in Figure 9. In summary, xPath achieved F1 scores of 0.755, 0.904, 0.763, and 0.946 in identifying the pathological patterns of mitosis, necrosis, prominent nucleolus, and sheeting. The scores indicate the effectiveness of our models. Moreover, for the tasks of cell-counting in hypercellularity and Ki-67 proliferation index criteria, we test their performance with 150 randomly-selected $512 \times 512 \times 3$ tiles each and report the average error rate. The results show that the average error rate of nuclei counting (hypercellularity) and Ki-67 proliferation index is 12.08% and 29.36%, respectively.

Dataset	Dimension (in pixels)	# of Samples (Training)	# of Samples (Testing)
Mitosis	256 × 256 × 3	33,562 (1,925 positive, 31,637 negative)	8,223 (336 positive, 7,887 negative)
Necrosis	512 × 512 × 3	4,383 (from 190 regions) (651 positive, 3,732 negative)	3,587 (from 162 regions) (770 positive, 2,817 negative)
Prominent Nucleoli	96 × 96 × 3	15,042 (2,447 positive, 12,595 negative)	3,753 (609 positive, 3,144 negative)
Sheeting	240 × 240 × 3	3,660 (from 55 regions) (1605 positive, 2055 negative)	2,340 (from 45 regions) (1,185 positive, 1,155 negative)

Table 2. The description of the dataset for each task. The dimensions of input tiles (in pixels), the size of training/testing sets, and the distribution of positive/negative tiles are provided.

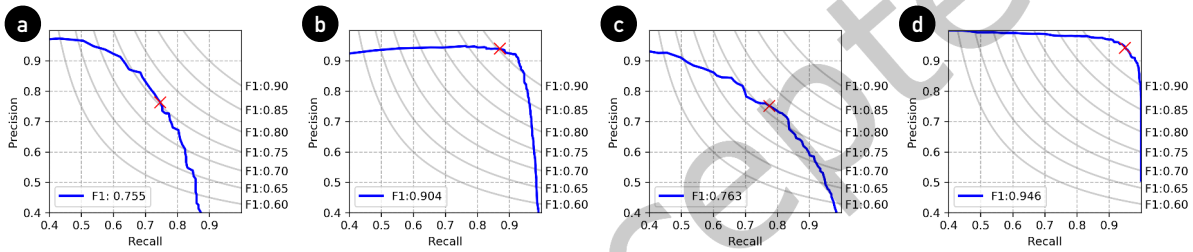


Fig. 9. Classification performance for (a) mitosis, (b) necrosis, (c) prominent nucleoli, (d) sheeting. The solid blue lines in each sub-figure illustrate the Precision-Recall curves of each model. The red crosses indicate the performance achieved by the models using the thresholds that maximized the F1 scores on the validation sets. The gray lines in each figure are the height lines of the F1 scores. The F1 score of each height line is shown on the right axis.

Due to a lack of data at present, for brain invasion and small cell patterns, rather than drawing a definitive conclusion, xPath uses a rule-based, unsupervised approach to recommend areas for pathologists to examine. We planned to validate the performance on these two criteria later in the work sessions with pathologists; however, it was hard for the participants to differentiate the small cell formation *vs.* inflammation areas without proper IHC tests. As such, xPath's AI performance in detecting small cell patterns was not validated. For the brain invasion, most pathologists felt it was faster to examine it manually and did not rely on AI's recommendations.

7 WORK SESSIONS WITH PATHOLOGISTS

The technical evaluation reported in the previous session validated the effectiveness of xPath's AI backend in the in-house dataset. However, it remains unanswered whether xPath is beneficial to pathologist users in practice. Notably, many previous cases showed how easily AI models could break, although they showed high accuracy in training/test data [47, 74]. To address these concerns, we conducted work sessions with 12 medical professionals in pathology across three medical centers and studied their behavior of grading meningiomas using a traditional interface — an open-source whole slide image viewer called ASAP[7] and xPath. In this study, we referred to

[7]https://computationalpathologygroup.github.io/ASAP/. This tool was selected because it is open-source and has gained popularity in the digital pathology research domain [54].

the traditional interface as system 1 and xPath as system 2 to avoid biasing of participants. The main research questions are:

RQ1: Can xPath *enable pathologists to achieve accurate diagnoses?*

One reason for utilizing AI in xPath is because it can highlight ROIs of multiple pathological patterns, freeing pathologists from examining the entire slide. However, it is still yet unclear whether introducing AI will have a positive or a negative effect on pathologists' diagnoses: On one hand, multiple previous works show that the introduction of human-AI collaboration improves pathologists' performance [15, 83]; On the other hand, due to the existing limitations in AI models' accuracy, users face the risk to generate wrong diagnoses if they over-rely on the non-perfect AI [7, 14]. Since there is no solid conclusion on this, we hypothesize that —

- **[H1] Pathologists' grading decisions with** xPath **will be as accurate as those with manual examinations.**

RQ2: Do pathologists work more efficiently with xPath*?*

Another reason for using AI in xPath is that it can improve the pathologists' throughput by alleviating their workload. However, it remains unanswered how AI will assist pathologists in xPath, given that previous work shows less-carefully-designed AI might incur extra burdens [35]. As such, it is also necessary to find out whether pathologists can work efficiently with xPath's AI. We hypothesize that —

- **[H2a] Pathologists will spend less time examining meningioma cases using** xPath**.**
- **[H2b] Pathologists will perceive less effort using** xPath**.**

RQ3: Overall, does xPath *add value to pathologists' existing workflow?*

Going beyond the influence brought by AI, we introduce two design ingredients for pathology AI systems — joint-analyses of multiple criteria *and* explanation by hierarchically traceable evidence in xPath. We also concluded three system requirements, *i.e.*, comprehensiveness, explainability, and integrability for xPath. In this study, we investigate whether such designs will add value to pathologists' existing workflow. Specifically, we hypothesize that:

- **[H3a]** xPath **will improve comprehensiveness with the joint-analyses of multiple criteria.**
- **[H3b]** xPath **will improve explainability with explanation by hierarchically traceable evidence.**
- **[H3c]** xPath **will improve integrability with the top-down human-AI collaboration workflow.**

7.1 Participants

We recruited 12 medical professionals in pathology across three medical centers in the United States through word-of-mouth and by sending flyers to the mailing lists. All participants were required to complete at least one year of post-graduate pathology residency training (≥ PGY-2). Our participants' experience ranged from two to ten years (μ=4.38, σ=2.16), including two attendings (A), two fellows (F), seven senior residents (SR, ≥ PGY-3), and one junior resident (JR, PGY-2). The demographic information of the participants is shown in Table 3. All participants had received training for examining meningiomas before the work sessions. And all participants had experience in seeing digital pathology slides prior to the study. They primarily used the Imagescope (a commercial software that provides image viewing functions similar to the ASAP) to see whole slide images (WSIs). The primary purpose of using the digital system was to train or review remote cases.

7.2 Test Data

We asked our pathologist collaborators in a local medical center to select 18 meningioma slides and scan them to WSIs[8] with an Aperio CS2 scanner to generate the test cases (IRB#20-000431). In normal conditions, each patient's case consisted of more than 10 WSIs, and an averaged-experienced resident pathologist typically needs

[8]…which include eleven H&E WSIs (scanned in x400), and seven Ki-67 WSIs (scanned in x200).

ID	Occupation	Years of Experience	Frequency of Seeing WSIs	ME194	ME195	ME196	ME197	ME198	ME199
P1	PGY-3	3	Weekly		ASAP	xPath			
P2	PGY-4	4	Monthly	ASAP		xPath	xPath	xPath	xPath
P3	Fellow	4	In Six Months			xPath		ASAP	
P4	Fellow	5	Weekly		xPath	ASAP		xPath	
P5	PGY-4	4	Weekly	xPath	xPath			ASAP	
P6	PGY-3	3	Monthly			xPath	ASAP		
P7	Attending	7	Weekly				xPath	ASAP	xPath
P8	PGY-4	3.5	Weekly		xPath				ASAP
P9	PGY-2	2	Bi-weekly	ASAP			xPath		
P10	PGY-3	3	Weekly	xPath			ASAP	xPath	
P11	PGY-4	4	Monthly		ASAP	xPath			
P12	Attending	10	Weekly		xPath	xPath		ASAP	

Table 3. Demographic information & arrangements of the participants in the work sessions. 'ME195' – 'ME199' are the case IDs. During the study, participants used 'ASAP' (system 1) and 'xPath' (system 2) to examine the cases. Note that FP12 had also participated in the formative study (referred to as FP3 in Table 1.)

to spend about one hour to finish examining an averaged-difficult case (*i.e.*, criteria found in the case do not lie on the grading borderlines). As such, we generated nine 'virtual patient cases' with the 'virtual cookie cut' technique (see Figure 10) to fit the task of grading meningiomas in hour-long working sessions.

Each virtual patient consisted of a mandatory H&E slide (in x400), and an optional Ki-67 slide (in x200). Each H&E slide had two nodes (each has a size of 30,000×30,000 pixels), while each Ki-67 slide had two corresponding Ki-67 nodes (each has a size of 15,000×15,000 pixels) that were extracted from the same position as their H&E counterparts, if available. The contours of nodes were removed as a "wash-out" measure because some participants had seen the slides before the study. All nodes were selected by an expert pathologist and included deterministic regions of interest (*i.e.*, crucial areas that include necessary information) for the diagnosis. Therefore, although participants were seeing virtual patients in the study, they still had to use the full system to diagnose because pathological criteria in the test data were not eliminated. In total, nine virtual cases have nine H&E slides and six Ki-67 slides.

The ground truth diagnoses was provided by an experienced pathologist, including two WHO grade 1, five WHO grade 2, and two WHO grade 3. We selected three from the grade 2 cases for the tutorial purpose, leaving the test set with two cases for each grade.

7.3 Task & Procedure

All sessions were conducted online because of the COVID-19 pandemic. We first introduced the project's mission and provided a detailed walkthrough of the traditional interface and xPath with three pairs of H&E and Ki-67 slides as an example. Participants used Microsoft Remote Desktop to interact with both systems that ran on a remote server. Next, we ran a testing session for the participants to grade one virtual case with the traditional interface, and one-four others using xPath with the time cost logged. The variation in the cases was caused by the between-subject difference in the time consumption of using xPath. And such a difference was caused by

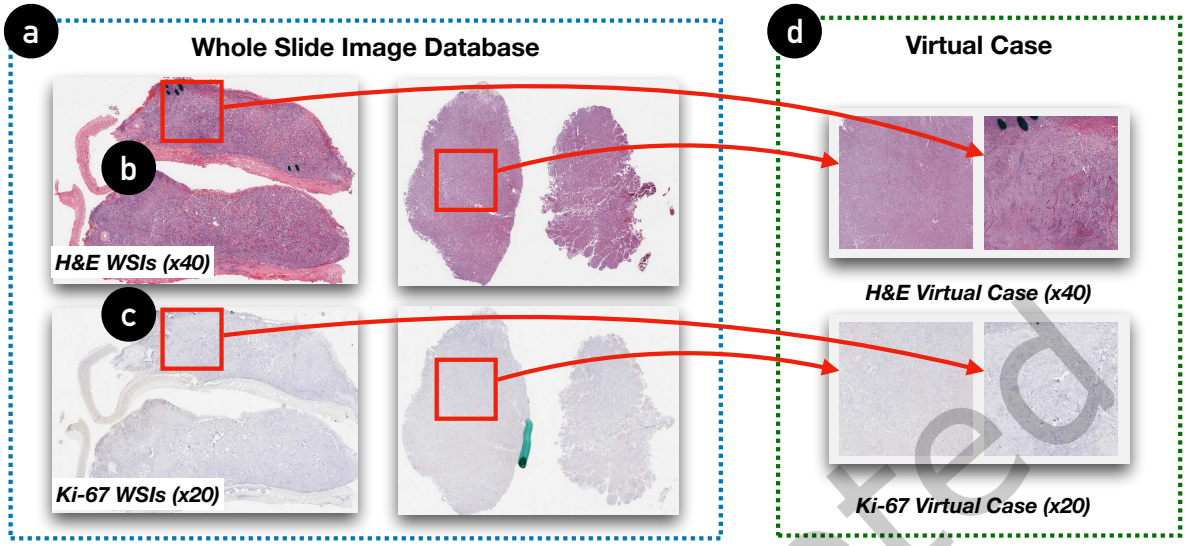


Fig. 10. We used the 'virtual cookie cut' technique to generate the tests cases. Specifically, we first collected (a) pairs of H&E (in x400) and Ki-67 (in x200) WSIs. Then, we generated 'virtual cuts' by (b) selecting 30,000×30,000-pixel regions in H&E WSIs, and (c) 15,000×15,000-pixel regions from the same position as their H&E counterparts. (d) Each virtual case consists of one mandatory H&E slide with two nodes and one optional Ki-67 slide with two corresponding ones.

two factors: *(i)* participants' learning abilities — some learned faster to use xPath than others; *(ii)* participants' abilities in examining the evidence. The order was counterbalanced across participants.

For each case, the time was counted from when participants first clicked the WSI case until they reached the grading diagnosis. After participants finished each case, we asked them to report their grading diagnosis as well as their findings through a questionnaire adapted from the College of American Pathologists (CAP) cancer protocol template[9]. In this session, we did not compare xPath with traditional optical microscopes because of the difficulty of instrumentation and observation given the remote situation. After participants had examined all the cases, we conducted a semi-structured interview to elicit their responses to xPath's perceived effort and added value. The average duration of each work session was ~70 minutes. Although conducted online, we set up the testing environment as close to pathologists' everyday clinical workflow: *(i)* we used H&E and Ki-67 data based on real patients (as described in Section 7.2); *(ii)* we used real working systems of ASAP and xPath; *(iii)* we asked our participants to diagnose following the same examination protocol as they had done in practice.

7.4 Measurements

In this study, we collected participants' grading decisions from the CAP questionnaire and analyzed the time log. We also asked them to fill in a post-study questionnaire (see Table 4) with seven-point Likert questions following [16, 37, 46]. We tested our hypotheses via the following measurements:

For **H1**, we compared the diagnoses reported by participants and the ground-truth diagnoses. We measured the accuracy of both systems by calculating the error rates.

[9]https://documents.cap.org/protocols/cp-cns-18protocol-4000.pdf

For **H2a**, we calculated the average time participants spent on each case using xPath and the traditional interface. For **H2b**, we asked them to give both systems ratings of the effort needed for grading (Table 4, W1), and the effectiveness of the system in reducing the workload (Table 4, W2) in the post-study questionnaire.

H3a-c was evaluated by the post-study questionnaire. For **H3a**, we asked participants to rate the comprehensiveness of xPath and the traditional interface (Table 4, C1). For **H3b**, we asked them to rate the explainability of xPath only since the traditional interface did not provide AI detections (Table 4, E1). For **H3c**, we asked participants to rate the integrability of both systems (Table 4, I1). Because "comprehensiveness", "explainability" and "integrability" are non-trivial terms, we included the following clarifications for the three terms in the questionnaire:

- **"Comprehensiveness"**: "*whether the system can provide detections for (1) multiple criteria for diagnosis and (2) entire slide, instead of a local area;*"
- **"Explainability"**: "*(1) how results from multiple criteria are combined to yield a grading; (2) what evidence leads to the value of each criterion; (3) why AI thinks a piece of evidence is positive / negative;*"
- **"Integrability"**: "*whether the system is integrable to your workflow of examining meningiomas.*"

Apart from the hypotheses, we also asked the participants to rate the helpfulness of each component in xPath ("*Rate the helpfulness of each component.*" — 1=lowest and 7=highest). Next, we investigated whether the participants trusted xPath by asking them the following two questions: *(i) How capable is the system at helping grade meningiomas?* (Table 4, T1), *(ii) How confident do you feel about the accuracy of your diagnoses using the system?* (Table 4, T2). Last but not least, to evaluate participants' attitudes towards xPath's workflow integration, we asked whether the participants would like to use both systems in the future (Table 4, F1), and also let the participants rate the overall preference of system 1 *vs.* system 2 (Table 4, F2).

8 RESULTS & FINDINGS

In this section, we first discuss our initial research questions and hypotheses. Then, we summarize the recurring themes that we have found in the working sessions.

8.1 RQ1: Can xPath enable pathologists to achieve accurate diagnoses?

We summarize the CAP questionnaire responses from our participants and collect 12 grading decisions from the traditional interface and 20 from xPath. We then follow previous works on digital pathology [73, 78] and compare the difference between participants' responses and the ground truth diagnoses. In summary, with the traditional interface, participants gave correct grading decisions for 7/12 cases, lower-than-ground-truth gradings for 4/12 cases, and higher-than-ground-truth grading for 1/12 cases. In comparison, using xPath, participants gave 17/20 cases correct gradings and lower-than-ground-truth gradings in 3/20 diagnoses. Upon further analysis, we found that all three errors that participants made with xPath were caused by their over-reliance on AI. In these cases specifically, participants spent the majority of their effort examining the evidence reported by xPath and missed the false-negative features that xPath failed to detect —

> *It's just that I got caught up in looking at the boxes, and I would forget that I should look at the entire case myself.* (P4)

In sum, based on the data collected by the study, we report that participants could make more accurate grading decisions with xPath compared to the traditional interface (**H1**).

8.2 RQ2: Do pathologists work more efficiently with xPath?

Contrary to our hypothesis (**H2a**), participants spent an average of 7min13s examining each case using xPath, which is 1min17s higher than the traditional interface (ASAP). Our study suggests that participants tended to (p=0.050, Wilcoxon rank-sum test, same below) invest more time in xPath than the traditional interface. We

Questions	ASAP	xPath
C1: Rate the comprehensiveness of the system.	2.83(1.27)	5.75(0.75)
E1: Rate the explainability of the system.	N/A	5.58(0.90)
I1: Rate the integrability of the system.	4.17(1.70)	5.91(1.08)
W1: Rate the effort needed to grade meningiomas when using the system.	3.67(1.37)	0.91(0.90)
W2: Rate the effect of the system on your workload to reach a diagnosis.	2.17(1.40)	5.83(1.03)
T1: How capable is the system at helping grade meningiomas?	N/A	5.83(0.94)
T2: How confident do you feel about the accuracy of your diagnoses using the system?	N/A	6.00 (0.95)
F1: If approved by the FDA, I would like to use this system in the future.	3.75(1.76)	6.42(0.79)
F2: Overall preference	6.75(0.45)	

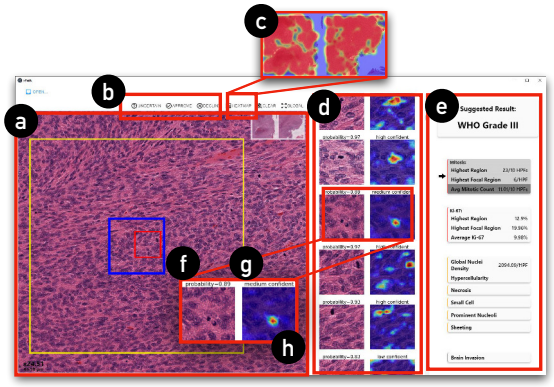
Table 4. Participants' response of average scores (and standard deviation) on the quantitative measurements of a traditional interface (ASAP) and xPath with seven-point Likert questions. For the rating questions (C1, E1, I1, W1, W2), 1=lowest and 7=highest. For question T1, T2, F1, 1=very strongly disagree, 2=strongly disagree, 3=slightly disagree, 4=neutral, . . . , and 7=very strongly agree. For question F2, 1=totally prefer system 1 over system 2, 2=much more prefer system 1 over system 2, 3=slightly prefer system 1 over system 2, 4=neutral, . . . , and 7=totally prefer system 2 over system 1. Note that for question W1, a higher score indicates that users perceive more effort while using the system. Question E1, T1, T2 are not applicable to ASAP, since it does not provide AI assistance.

believe this is partly because xPath brings participants an extra workload to comprehend and justify the AI findings. In the traditional interface, our participants share a similar workflow of examining the WSI — they first scanned the entire WSI in low magnification, then prioritized studying one criterion (such as the brain invasion or the mitotic count) to ascertain a probable diagnosis as quickly as possible. They also checked Ki-67 slides to support their diagnosis. In this process, they collected evidence that accounts for a higher grade and memorized them in their minds. Once they acquired enough evidence, they would stop and make a grading decision. When using xPath, participants did not abandon their standard workflow as in the traditional interface. Rather, on top of their standard workflow, participants would perform the differential diagnosis based on AI's findings — they clicked through each piece of evidence in xPath, justified it by registering into the WSI, and at times overrode AI by clicking the approve/decline/declare-uncertain buttons. These extra steps of interactions prolong participants' workflow —

> *System 2 (*xPath*) actually makes it longer because some of the images have sort of competing opinions — whether this is mitosis or not . . . So I'd better take a closer look at what the machine suggests.* (P3)

Regarding the perceived effort (**H2b**), participants reported significantly less effort (Table 4, W1, xPath: μ=0.91, ASAP: μ=3.67, p=0.002) and a stronger effect on reducing the workload (Table 4, W2, xPath: μ=5.83, ASAP: μ=2.17, p=0.002) while using xPath. Participants mentioned that automating the process of finding small-scaled histopathological features, especially mitosis, would save their time and effort —

> *I spend a lot more time crawling around the slide in the high-power, looking for mitosis (for system 1), which you don't have to do as much in system 2 (*xPath*).* (P8)



		Rate the helpfulness of each component: (1: lowest → 7: highest)							
		1	2	3	4	5	6	7	Mean (Std)
System Component	WSI Viewer (a)	0	0	0	2	0	5	5	6.08 (1.08)
	Approve/Decline/ Uncertain (b)	1	0	0	1	0	3	7	6.00 (1.81)
	Heatmap (c)	0	0	0	2	3	5	2	5.58 (1.00)
	List of Sampled Evidence (d)	0	1	1	0	0	3	7	6.00 (1.71)
	List of Multiple Criteria (e)	0	0	0	0	2	2	8	6.50 (0.80)
Explainable Evidence	Probability (f)	3	1	1	2	1	3	1	3.83 (2.20)
	Confidence Level (g)	3	1	0	1	4	3	0	3.92 (2.07)
	Saliency Map (h)	0	1	0	3	2	4	2	5.17 (1.47)

Fig. 11. Participants' helpfulness ratings of each component in xPath. Each letter-labeled component in the right table corresponds to the marked part on the left.

8.3 RQ3: Overall, does xPath add value to pathologists' existing workflow?

For the comprehensiveness dimension (**H3a**), xPath received a significantly higher rating than the traditional interface (Table 4, C1, xPath: μ=5.75, ASAP: μ=2.83, p=0.001). Furthermore, participants gave an average helpfulness score of 6.50/7 for the design of joint-analyses of multiple criteria (see Figure 11e). They responded positively that such a design provides sufficient information (*i.e.*, criteria and evidence) to assist the diagnosis —

> *…it (xPath) kind of gives you a step-wise checklist to make sure that it's the correct diagnosis, and also provides you what is most likely a diagnosis.* (P11)

For the explainability dimension (**H3b**), xPath obtained an average rating of 5.58/7 (Table 4, E1). In general, participants could understand the logical relationship between the evidence and the suggested grading (global explainability). They also gave a high helpfulness rating (6.00/7, Figure 11d) for the list of evidence provided by xPath. However, participants gave lower ratings on the probability (3.83/7, Figure 11f) and the confidence level (3.92/7, Figure 11g) elements in the mid-level samples because they were hard to read in xPath —

> *"…these small words (pointing to the probability) …I didn't notice that very much …also it wasn't very easy to see."* (P3)

The saliency map received a relatively higher rating (5.17/7, Figure 11h). However, some (P1, P5) participants found it hard to interpret the saliency map, especially for the cases where cues of attention were scattered across the entire evidence (see Figure 13a) —

> *For the heatmap (the saliency map) …it is also a little bit confusing …it takes some time getting used to it and there are some false positives.* (P1)

For the integrability dimension (**H3c**), participants gave overall higher scores for xPath (Table 4, I1, xPath: μ=5.91, ASAP: μ=4.17, p=0.006). Specifically, participants were able to perform diagnoses based on the xPath's AI findings, which is similar to their workflow of collaborating with human trainees —

> *It's kind of like a first-year resident marking everything.* (P1)
>
> *I'm a cytology fellow, and cases are pre-screened for us. And essentially this is doing similarly.* (P4)

For the trust dimension, participants responded positively to xPath's capability of helping to grade meningiomas (T1: μ=5.83) and their accuracy of the diagnoses while using the system (T2: μ=6.00). However, some (P3,

P4, P5) pointed out that they would spend more time examining the WSI entirely if more time had been granted —

> *I just went to the areas that the system suggested. If I had more time, I would like to just go to all the areas, just to feel more comfortable that I'm not missing anything.* (P5)

Last, participants were more likely to use xPath than the traditional interface (Table 4, F1, xPath: μ=6.42, ASAP: μ=3.75, p=0.002). Overall, 9/12 of the participants "totally" preferred xPath over the traditional interface, while 3/12 "much more" preferred xPath (Table 4, F2).

However, it is noteworthy that this study is based on participants' examination of WSIs, while pathologists use the optical microscope in their daily practice. During the study, 7/12 of our participants expressed that they preferred using an optical microscope with the glass slide *vs.* a digital interface with the WSI — *"… it's much faster (in the microscope) than moving on the computer … we would prefer to look at a real slide instead of using a scan picture."* (P2). As such, further comparison between xPath and the optical microscope is considered future work.

8.4 Recurring Themes

We analyzed the video recordings of the work sessions in a similar approach as described in Section 4.1. Based on our observations of participants' using xPath and the interview with them, we discuss the following recurring themes that characterize how participants interacted with xPath.

8.4.1 Pathologists examine xPath*'s multiple criteria findings by prioritizing one and referring to others on demand.* We noted that participants tended to focus on a specific criterion. If that criterion alone did not meet the bar of a diagnosis for a higher grade, participants would use xPath to browse other criteria, looking for evidence of a differential diagnosis, until they identify sufficient evidence to support their hypothesis.

> *I'm done. Because with the mitosis that high, you're done. You don't have to go through that stuff (other criteria).* (P12)

However, some participants would also like to see other criteria and examine the slide comprehensively —

> *With the mitosis rate that high, you don't actually need it (Ki-67) for the diagnosis. But I will have a look at it.* (P1)

> *I will just look at (other criteria) because I don't want to grade by one single criterion (mitosis).* (P3)

Such a relationship between criteria is analogous to 'focus + context' [18] in information visualization — different pathologists might focus on a few different criteria. Still, the other criteria are also important to serve as context at their disposal to support an existing diagnosis or find an alternative.

8.4.2 xPath*'s top-down workflow with hierarchical explainable evidence enables pathologists to navigate between high-level AI results & low-level WSI details.* One of the main reasons limiting the throughput of histopathological diagnosis is that criteria like mitotic count have very small size compared to the dimensions of WSIs. As a result, participants have to switch to high magnification to examine such small features in detail. Given the high resolution of the WSI, it is possible to 'get lost' in the narrow scope of HPF, resulting in a time-consuming process to go through the entire WSI. With xPath, participants found its hierarchical design and the provision of mid-level evidence (*e.g.*, AI's ROI samples) the most helpful for diagnosis as it connects high-level findings and low-level details —

> *It (*xPath*) finds the best area to look at. … You can jump there, and if it is a grade 3, then it is a grade 3. You don't have to look at other areas.* (P6)

Furthermore, participants appreciated that xPath provided heatmap visualizations to assist them in navigating the WSI out of the ROI samples —

> *The heatmap is very useful to assist pathologists to go through the entire slide … which saves time and makes sure not missing anything.* (P12)

8.4.3 xPath*'s explainable design helps pathologists see what AI is doing.* We found xPath's evidence-based justification of AI findings assisted participants in relating AI-computed results with evidence, which added explainability —

> *System 2 (*xPath*) does find some evidence and assigns it to a particular observation that is related to the grading, so that it helps with explainability.* (P3)

In xPath, the AI might make two types of mistakes that may incur potential bias: *(i)* false positive, where AI mistakenly identifies negative areas as positive for a given criterion; *(ii)* false negative, where the AI misses positive areas corresponding to a criterion. We observed a number of false-positive detections that confused some participants. We also found out that the participants would rather deal with more false positives than false negatives so that signs of more severe grades would not be missed —

> *It's better that it picks them up and gives me the opportunity to decline it.* (P10)

Furthermore, although some participants found the saliency map hard to interpret in some cases, others used it to locate the cells that led to AI's grading —

> *There were a couple of instances where it was a bit more difficult to figure out what it (the saliency map) was trying to point out to me. But for the majority of the time, I could tell which area they (the saliency maps) were trying to show me.* (P9)

Further, with the aid of the saliency map, participants could understand AI's limitations and what might have misled the AI —

> *You can see what this system counted as mitosis … the heatmap (the saliency map) helps to understand why AI chose this or that area. For example, I think AI chose neutrophils as mitotic figures in some areas.* (P6)

8.4.4 Pathologists justify xPath *by incrementing human findings onto justified AI results.* Given the explainable evidence provided by xPath, it was straightforward for participants to recognize and modify AI results when there was a disagreement. Specifically, participants could justify AI by clicking on the approve/decline/declare-uncertain buttons or modifying AI results directly on the criteria panel. If the justified AI results were sufficient to conclude a grading decision (*e.g.*, seven mitoses in 10 HPFs, enough to make the case as grade 2 (>4), but still far from grade 3 («20)), they would stop examining and report the grading. However, if the justified AI results appeared to be marginal (*e.g.*, 19 mitoses in 10 HPFs, which is only one mitosis away from upgrading the case to a grade 3), participants would continue to search based on the AI findings and add their new insights to grade —

> *I count a total number of five … adding the previous 19 makes it 24 … this is grade 3.* (P2)

What's more, for the cases where xPath did not actively report positive detections, participants would examine the WSI manually as in a traditional interface — that is, participants would use their experience to evaluate the case further and make a grading decision.

9 DISCUSSION

In this section, we start by discussing this work's limitations and potential future improvements. We then summarize the design recommendations for future physician-AI collaborative systems. Finally, we focus on future directions for improving AI's integration into pathologists' workflow.

9.1 Limitations & Future Improvements

We conclude the following limitations of our current work:

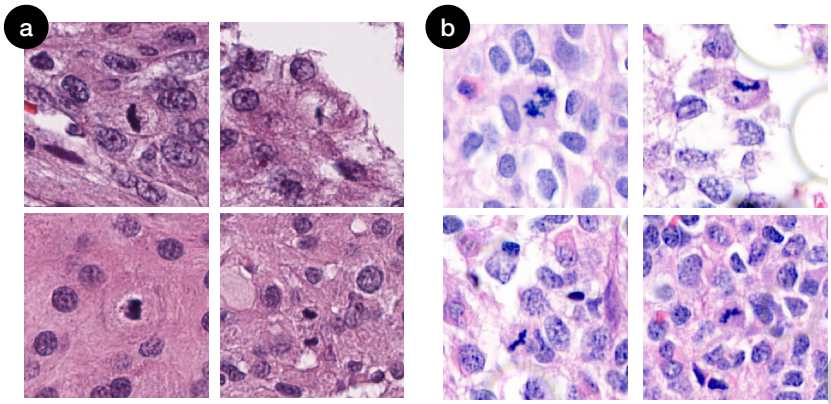


Fig. 12. Mitoses from meningiomas (in x400), scanned by (a) the medical center in this study and (b) a different medical center. The difference in appearance is caused by the difference in processing procedures and scanners used.

- xPath was evaluated on a small number of participants examining limited materials using a remote setup. As such, the observations and conclusions are inevitably biased and speculative;
- The AI's testing performance in this work was reported from an in-house dataset that was collected from one institute, while the evaluation of AI's alignment with the benchmarks from a large set of images from multiple medical centers was not conducted;
- xPath currently does not support users to adjust the cut-off prediction threshold, hence resulting in an amount of false-positive evidence;
- Cases of the saliency map (see Figure 13) confuse some participants because they can not highlight cells appropriately;

Next, we will discuss the limitations and future improvements in detail.

9.1.1 Increasing the scope of xPath*'s evaluation study.* The scope of xPath's evaluation study was limited to the following four aspects:

Study Material. Due to the Institutional Review Board (IRB) regulations, only a limited number of images from one medical center were selected and used in xPath. This leaves the performance of xPath's AI questionable while being applied to images from other institutes. This is because other institutes might use a different staining process or a different type of scanner, causing a difference in the image domain/distribution (see Figure 12). Furthermore, the limited test cases generated for xPath's work sessions might not reflect the distributions of meningiomas in clinical settings.

Participants: Recruitment and Sampling. Because of the rare availability of medical professionals in neuropathology, we only recruited twelve participants for the study, most of whom were residents. This might cause the conclusions for **RQ1** and **RQ2** inevitably speculative because research has shown that pathologists' diagnostic accuracy might be related to their experience level [33]. Moreover, all participants came from one country, which might cause the qualitative observations to be biased since no pathologists from other countries were involved.

Study Set-Up. All studies were conducted online due to the COVID-19 pandemic. And the duration of each study (about 60 minutes) was relatively short in order to prove the long-term validity of xPath. Additionally, no clinical testing was conducted because of strict legislation regulations from US Food and Drug Administration (FDA).

Apparatus. The comparison between the xPath and the optical microscope — pathologists' first approach to seeing pathology slides, was not conducted. Although the FDA has lifted its restrictions on digital whole slide images for clinical use since 2017 [29], we found it is still challenging to persuade pathologists to move from the optical microscope to the digital interfaces (without AI): more than half of the participants expressed that they preferred using an optical microscope with the glass slide *vs.* a traditional digital interface. Remarkably, participants found it challenging to navigate a digital whole slide image, which has also been described and discussed by Ruddle *et al.* [69]. However, our study found that pathologists preferred to use xPath because it adds value to their workflow with AI. Therefore, we suggest that future medical systems highlight their benefit to pathologists as an incentive to overcome the limitations in traditional digital interfaces.

In sum, future works should consider using more images from multiple medical centers, recruiting more participants with multiple experience levels, conducting long-term, in-person studies, and comparing xPath with the optical microscope. With more data points collected, we can validate xPath's performance and generalizability more comprehensively.

9.1.2 Enabling adjusting the thresholds within the interface. Currently, xPath does not support directly changing the threshold for a positive result with the interface. In our user study, one participant mentioned that different pathologist might have different thresholds to call whether a piece of evidence is positive —

> *"I only call the characteristic mitoses . . . other pathologists might have different thresholds.* (P7)

Further, dealing with false positives and false negatives is another issue with the fixed-threshold scheme. From our study, we found out that pathologists would prefer high-sensitivity results that include some false positives rather than high-specificity results that have false negatives —

> *I could have more faith if it could find all the candidates. And I could pretty easily click through and accept/reject, and know that it wasn't missing anything.* (P8)

Therefore, the system, by default, should be designed to err on the side of caution, *e.g.*, showing a wide range of ROIs despite some being inevitably false positives. Pathologists are fast in examining ROIs (and ruling out false positives), whereas missing important features would come with a much higher cost (*e.g.*, delayed or missed treatment).

9.1.3 Improving the quality and granularity of explanations. In the study, we found a number of cases where the saliency maps failed to explain the classification predictions and caused confusion to the users. As shown in Figure 13, the failed saliency maps showed either scattered attention across the evidence (Figure 13a), or concentrated attention at the wrong place (Figure 13b). Such errors can be explained as the attention is reasoned from patch-wise annotations rather than localized ones because the localized annotations of positive findings are extremely labor-costly to obtain. The quality of the saliency maps can be potentially improved with the increment of training data for higher model generalization and the advent of the methodologies of unsupervised attention reasoning [4].

Besides, knowing the location of a potential positive finding can be insufficient for pathologists. Since the pathological imaging of tissues is merely an approximation of the real condition, there can often exist uncertainty in diagnosis even for well-trained pathologists. As such, explaining why an area contains positive findings, *e.g.*, a highlighted cell is detected to stage as mitosis since its boundary is jagged, can be critical for systems in the future. Such causality enables a system to imitate how pathologists discuss with their peers, which can improve the collaboration between a system and its users. Moreover, future work should also employ more formal measurements (*e.g.*, System Causability Scale [40]) to evaluate the quality of explanations.

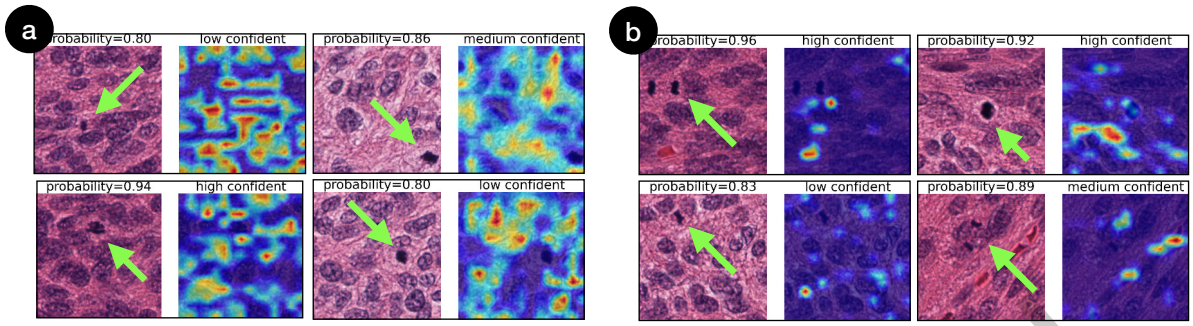


Fig. 13. Examples of failure explanation cases, where the saliency shows (a) scattered attention across the image or (b) misleading hot spots. The green arrows point to the location of a mitosis figure marked by a human pathologist.

9.2 Design Recommendations for Physician-AI Collaborative Systems

Although we focus on the grading of meningiomas in this work, we believe our two designs in xPath — joint-analyses of multiple criteria and explanation by hierarchically traceable evidence — can be generalizable to other medical applications that require doctors to see and verify numerous criteria from various medical tests (such as grading astrocytoma, IDH mutant (WHO Grade 2-4), solitary fibrous tumor (WHO Grade 1-3) [58]). Here, we provide design recommendations for future physician-AI systems.

9.2.1 Showing the logical relationships amongst multiple types of evidence at the top level. Carcinoma grading usually involves examining multiple criteria from various data sources (*e.g.*, H&E slides, IHC test, FISH (fluorescence in situ hybridization) test, patient's health record). As such, one-size-fits-all AI models are not sufficient. In practice, multiple AI models are employed to locate different types of disease markers. To organize these AI-computed results, medical AI systems (such as xPath) should seek to present the logical relationship that connects these multiple criteria/features/sources of information and update final results dynamically given any pathologists' input (*e.g.*, acceptance or rejection of how AI computes each criterion). Such a design is more likely to match the clinical practice of pathologists and cost minimal extra learning when users onboard a system.

9.2.2 Making AI's findings traceable with hierarchically organized evidence. There is a pressing need to deal with the transparency of a black-box model and the traceability of the explanation evidence in high-stakes tasks (*e.g.*, medical diagnosis). As such, AI systems should provide local explainability where each piece of low-level evidence is traceable. In xPath, we employ the design of hierarchically traceable evidence for each criterion. Such an organization forms an 'evidence chain' where each direct evidence is accountable for the high-level system output. Similar intuitions can also be applied to medical applications in a more general context, such as cancer staging [59] and cancer scoring [42], where the evidence is accumulated to arrive at a diagnosis.

9.2.3 Employing a "focus+context" design toward presenting and/or interacting with multiple criteria. Medical diagnosis involves accumulating evidence from multiple criteria — our study observed that pathologists started by focusing on one criterion while continuing to examine the others for a differential diagnosis. Thus, medical AI systems should make multiple criteria available, and support the navigation of such criteria following a "focus+context" design [18], which is commonly used in information visualization. The major design goal is to strike the dichotomy between juxtaposing the focused criterion with sufficient contextual criteria and overwhelming the pathologists with too much information. It is also possible for a system to, based on a patient's

prior history and the pre-processing of their data, recommend a pathologist to start focusing on specific criteria followed by examining some others as context.

9.3 On Integrating AI into Pathologists' Workflow

9.3.1 How has AI improved pathologists' diagnoses in xPath*?* Similar to previous human-AI collaborative research in medicine [30, 52], we discovered that using AI might improve pathologists' diagnosis quality. In pathology, the AI can "efficiently, systematically, exhaustively" analyze the entire whole slide image [73]. Therefore, xPath can help pathologist capture small-sized details they might miss in the manual examination, which can improve their sensitivity. xPath further aggregates these details into AI-recommended regions of interest (ROIs), and pathologists can check each ROI of each criterion. Compared to the manual examinations where pathologists have to see multiple criteria with one pass (*i.e.*, "multitasking", as described in Section 4), such a design assists less-experienced pathologists in examining in a more organized, more comprehensive manner.

Furthermore, xPath's ROI recommendations freed participants from heavy navigation and visual searching. Traditionally, pathologists navigate manually [63, 69] and search visually to locate pathological patterns. With xPath, our participants could see and adjudicate ROI recommendations directly. However, it is noteworthy that forcing pathologists to see ROI recommendations might break their workflow. First, because ROI recommendations are not necessarily physically adjacent, pathologists need to "jump" from one ROI to another to examine them. And it is unclear whether pathologists can accept such " ROI jumpings" without continuous navigation (*i.e.*, panning and zooming). Second, the presentation of ROI recommendations (*e.g.*, in xPath, boxes) may also influence pathologists' judgement — one participant expressed their concern when the ROI highlighted an area but failed to do so in a similar one — "*If I called this positive (pointing at one recommendation box), should I also call this one (pointing at another area but not marked by recommendation boxes)?*"(P7). Hence, we suggest that future HCI systems study pathologists' acceptance of using ROIs to examine and elaborate more on the over-reliance issues.

9.3.2 How to make human-AI systems in pathology more robust? Although incorporating AI might benefit users, the performance of human-AI collaboration workflow might be influenced in clinical settings [9, 84]. Therefore, it is crucial to design workflows that can cope with chaotic "in the wild" situations. xPath applied two designs to assist pathologists to debug and refine the AI findings: *(i)* hierarchical evidence that makes the AI analysis traceable and transparent; *(ii)* pathologists can refine the AI findings by approving/declining/declaring-uncertain AI analysis.

Based on the observations of how our participants interacted with xPath, we further discuss the potential approaches to make human-AI systems more robust for future pathology applications. The first approach is to add additional sources of information so that pathologists can verify the AI recommendations. For example, xPath mimics how pathologists examine meningiomas and adds an additional test — the Ki-67 test — for mitosis ROI recommendations. In our user study, we found that pathologists could cross-check the Ki-67 hot-spot areas with mitosis ROI recommendations to validate the correctness.

For the systems without the luxury of additional tests, we suggest re-framing the human-AI collaboration workflow by forcing doctors to give a brief overview first and then retrieve AI recommendations on demand. Such a strategy is called the "*cognitive forcing function*" and is viable for reducing the over-reliance issues in previous literature [14]. We argue that such a workflow design is still integrable to pathologists' practice because their manual examination also starts with an overview of a slide [69].

Finally, enabling users to control the recommendation process might also be a solution. For example, a slider can be used to control the sensitivities of AI-recommended ROIs. As such, pathologists can first see the most pressing ROI, and then gradually see more on demand. Such a design reduces the disruptive behavior of using AI systems in the wild and pathologists are more likely to accept it in practice [16].

9.3.3 How should AI systems build trust for pathologists? Previous HCI research advises informing doctors of AI's capabilities and limitations to gain trust [17]. For example, Sendak *et al.* created a "model fact sheet" inspired by pharmaceutical drug labels to inform doctors of AI details [72]. In our study, we also discovered that participants preferred to know the AI capabilities — "*I really wanna cross-check (AI's) accuracy with a human observer, and cases of a range of mitosis, from rare mitosis to frequent mitosis.*"(FP1) *"Pathologists are data-driven ... if you can show it (AI) is accurate for like 1,000 cases, they may buy it."*(P1) As such, we suggest future medical AI systems to demonstrate AI's capabilities by presenting with a set of examples with AI's predictions and ground truth. With the help of examples, pathologists can briefly evaluate AI performance and know its capabilities and limitations.

Apart from AI's information, previous studies indicate that explanations might improve trust: some attempt to explain AI predictions with XAI components (*e.g.*, the saliency map [91]), while others build inherently interpretable models (*e.g.*, concept bottleneck models [49]). During the study, we found that our participants preferred simple explanations during the interaction with xPath. Although complex explanations (*e.g.*, concept explanations) might provide a more detailed background, pathologists might justify a vast number of explanations during the time-pressing diagnosis process. If the explanations cannot capture pathologists' attention initially, they might ignore them for the rest of the examination process (also described by P3 in our user study). Therefore, we suggest future medical AI systems allow pathologists to see *levels of* explanations on demand. For example, pathologists might see simple visual explanations by default but can opt to see more detailed explanations if they wish.

10 CONCLUSION

In this work, we identify three challenges of comprehensiveness, explainability, and integrability that prevent AI from being adopted in a complex clinical setting for pathologists. To close these gaps, we implement xPath with two key design ingredients: *(i)* joint-analyses of multiple criteria and *(ii)* explanation by hierarchically traceable evidence. To validate xPath, we conducted work sessions with twelve medical professionals in pathology across three medical centers. Our findings suggest that xPath can leverage AI to reduce pathologists' cognitive workload for meningioma grading. Meanwhile, pathologists benefited from the design and made fewer mistakes with xPath, compared to the manual baseline interface. By observing pathologists' use of xPath and collecting their quantitative and qualitative feedback, we indicate how pathologists may collaborate with AI and summarize design recommendations. We believe that xPath is useful for other HCI research by providing first-hand information on how pathologists collaborate and manage multiple AI outcomes, which opens up a new space for pathologist-AI interaction possibilities.

ACKNOWLEDGMENTS

This work was funded in part by a the National Science Foundation under grant IIS-1850183. We appreciate the anonymous participants for the user study, and the editor and reviewers for their valuable feedback in improving the manuscript.

REFERENCES

[1] Ellen Abry, Ingrid Ø Thomassen, Øyvind O Salvesen, and Sverre H Torp. 2010. The significance of Ki-67/MIB-1 labeling index in human meningiomas: a literature study. *Pathology-Research and Practice* 206, 12 (2010), 810–815.

[2] Mohamed Amgad, Habiba Elfandy, Hagar Hussein, Lamees A Atteya, Mai AT Elsebaie, Lamia S Abo Elnasr, Rokia A Sakr, Hazem SE Salem, Ahmed F Ismail, Anas M Saad, et al. 2019. Structured crowdsourcing enables convolutional segmentation of histology images. *Bioinformatics* 35, 18 (2019), 3461–3467.

[3] Teresa Araújo, Guilherme Aresta, Eduardo Castro, José Rouco, Paulo Aguiar, Catarina Eloy, António Polónia, and Aurélio Campilho. 2017. Classification of breast cancer histology images using convolutional neural networks. *PloS one* 12, 6 (2017), e0177544.

[4] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies,

opportunities and challenges toward responsible AI. *Information Fusion* 58 (2020), 82–115.

[5] Eirini Arvaniti, Kim S Fricker, Michael Moret, Niels Rupp, Thomas Hermanns, Christian Fankhauser, Norbert Wey, Peter J Wild, Jan H Rueschoff, and Manfred Claassen. 2018. Automated Gleason grading of prostate cancer tissue microarrays via deep learning. *Scientific reports* 8, 1 (2018), 1–11.

[6] Thomas Backer-Grøndahl, Bjørnar H Moen, and Sverre H Torp. 2012. The histopathological spectrum of human meningiomas. *International journal of clinical and experimental pathology* 5, 3 (2012), 231.

[7] Gagan Bansal, Besmira Nushi, Ece Kamar, Walter S. Lasecki, Daniel S. Weld, and Eric Horvitz. 2019. Beyond Accuracy: The Role of Mental Models in Human-AI Team Performance. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 7, 1 (Oct. 2019), 2–11. https://ojs.aaai.org/index.php/HCOMP/article/view/5285

[8] Dalal Bardou, Kun Zhang, and Sayed Mohammad Ahmad. 2018. Classification of breast cancer based on histology images using convolutional neural networks. *IEEE Access* 6 (2018), 24680–24693.

[9] Emma Beede, Elizabeth Baylor, Fred Hersch, Anna Iurchenko, Lauren Wilcox, Paisan Ruamviboonsuk, and Laura M Vardoulakis. 2020. A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–12.

[10] Babak Ehteshami Bejnordi, Mitko Veta, Paul Johannes Van Diest, Bram Van Ginneken, Nico Karssemeijer, Geert Litjens, Jeroen AWM Van Der Laak, Meyke Hermsen, Quirine F Manson, Maschenka Balkenhol, et al. 2017. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *Jama* 318, 22 (2017), 2199–2210.

[11] Kaustav Bera, Kurt A Schalper, David L Rimm, Vamsidhar Velcheti, and Anant Madabhushi. 2019. Artificial intelligence in digital pathology—new tools for diagnosis and precision oncology. *Nature reviews Clinical oncology* 16, 11 (2019), 703–715.

[12] Christof A Bertram, Marc Aubreville, Taryn A Donovan, Alexander Bartel, Frauke Wilm, Christian Marzahl, Charles-Antoine Assenmacher, Kathrin Becker, Mark Bennett, Sarah Corner, et al. 2022. Computer-assisted mitotic count using a deep learning–based algorithm improves interobserver reproducibility and accuracy. *Veterinary pathology* 59, 2 (2022), 211–226.

[13] Daniel J Brat, Joseph E Parisi, Bette K Kleinschmidt-DeMasters, Anthony T Yachnis, Thomas J Montine, Philip J Boyer, Suzanne Z Powell, Richard A Prayson, and Roger E McLendon. 2008. Surgical neuropathology update: a review of changes introduced by the WHO classification of tumours of the central nervous system. *Archives of pathology & laboratory medicine* 132, 6 (2008), 993–1007.

[14] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z Gajos. 2021. To trust or to think: cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–21.

[15] Wouter Bulten, Maschenka Balkenhol, Jean-Joël Awoumou Belinga, Américo Brilhante, Aslı Çakır, Lars Egevad, Martin Eklund, Xavier Farré, Katerina Geronatsiou, Vincent Molinié, et al. 2021. Artificial intelligence assistance significantly improves Gleason grading of prostate biopsies by pathologists. *Modern Pathology* 34, 3 (2021), 660–671.

[16] Carrie J. Cai, Emily Reif, Narayan Hegde, Jason Hipp, Been Kim, Daniel Smilkov, Martin Wattenberg, Fernanda Viegas, Greg S. Corrado, Martin C. Stumpe, and Michael Terry. 2019. Human-Centered Tools for Coping with Imperfect Algorithms During Medical Decision-Making. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3290605.3300234

[17] Carrie J Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. 2019. " Hello AI": Uncovering the Onboarding Needs of Medical Practitioners for Human-AI Collaborative Decision-Making. *Proceedings of the ACM on Human-computer Interaction* 3, CSCW (2019), 1–24.

[18] Mackinlay Card. 1999. *Readings in information visualization: using vision to think*. Morgan Kaufmann.

[19] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. 2018. Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. 839–847. https://doi.org/10.1109/WACV.2018.00097

[20] Chhavi Chauhan and Rama R Gullapalli. 2021. Ethics of AI in pathology: Current paradigms and emerging issues. *The American Journal of Pathology* 191, 10 (2021), 1673–1683.

[21] Xiang 'Anthony' Chen, Ye Tao, Guanyun Wang, Runchang Kang, Tovi Grossman, Stelian Coros, and Scott E. Hudson. 2018. Forte: User-Driven Generative Design. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3173574.3174070

[22] Edward Choi, Mohammad Taha Bahadori, Joshua A. Kulas, Andy Schuetz, Walter F. Stewart, and Jimeng Sun. 2016. RETAIN: An Interpretable Predictive Model for Healthcare Using Reverse Time Attention Mechanism. In *Proceedings of the 30th International Conference on Neural Information Processing Systems* (Barcelona, Spain) *(NIPS'16)*. Curran Associates Inc., Red Hook, NY, USA, 3512–3520.

[23] Dan C. Cireşan, Alessandro Giusti, Luca M. Gambardella, and Jürgen Schmidhuber. 2013. Mitosis Detection in Breast Cancer Histology Images with Deep Neural Networks. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2013*, Kensaku Mori, Ichiro Sakuma, Yoshinobu Sato, Christian Barillot, and Nassir Navab (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 411–418.

[24] Alberto Corvò, Marc A. van Driel, and Michel A. Westenberg. 2017. PathoVA: A visual analytics tool for pathology diagnosis and reporting. In *2017 IEEE Workshop on Visual Analytics in Healthcare (VAHC)*. 77–83. https://doi.org/10.1109/VAHC.2017.8387544

[25] Pat Croskerry, Karen Cosby, Mark L Graber, and Hardeep Singh. 2017. *Diagnosis: Interpreting the shadows*. CRC Press.

[26] David Dov, Serge Assaad, Ameer Syedibrahim, Jonathan Bell, Jiaoti Huang, John Madden, Rex Bentley, Shannon McCall, Ricardo Henao, Lawrence Carin, and Wen-Chi Foo. 2021. A Hybrid Human–Machine Learning Approach for Screening Prostate Biopsies Can Improve Clinical Efficiency Without Compromising Diagnostic Accuracy. *Archives of Pathology & Laboratory Medicine* (09 2021). https://doi.org/10.5858/arpa.2020-0850-OA arXiv:https://meridian.allenpress.com/aplm/article-pdf/doi/10.5858/arpa.2020-0850-OA/2913491/10.5858_arpa.2020-0850-oa.pdf

[27] Mehmet Günhan Ertosun and Daniel L Rubin. 2015. Automated grading of gliomas using deep learning in digital pathology images: A modular approach with ensemble of convolutional neural networks. In *AMIA Annual Symposium Proceedings*, Vol. 2015. American Medical Informatics Association, 1899.

[28] Theodore Evans, Carl Orge Retzlaff, Christian Geißler, Michaela Kargl, Markus Plass, Heimo Müller, Tim-Rasmus Kiehl, Norman Zerbe, and Andreas Holzinger. 2022. The explainability paradox: Challenges for xAI in digital pathology. *Future Generation Computer Systems* 133 (2022), 281–296.

[29] Office of the FDA. [n. d.]. FDA allows marketing of first whole slide imaging system for Digital Pathology. https://www.fda.gov/news-events/press-announcements/fda-allows-marketing-first-whole-slide-imaging-system-digital-pathology

[30] Riccardo Fogliato, Shreya Chappidi, Matthew Lungren, Paul Fisher, Diane Wilson, Michael Fitzke, Mark Parkinson, Eric Horvitz, Kori Inkpen, and Besmira Nushi. 2022. Who Goes First? Influences of Human-AI Workflow on Decision Making in Clinical Imaging. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. 1362–1374.

[31] Marcel Gehrung, Mireia Crispin-Ortuzar, Adam G Berman, Maria O'Donovan, Rebecca C Fitzgerald, and Florian Markowetz. 2021. Triage-driven diagnosis of Barrett's esophagus for early detection of esophageal adenocarcinoma using deep learning. *Nature medicine* 27, 5 (2021), 833–841.

[32] Parmida Ghahremani, Yanyun Li, Arie Kaufman, Rami Vanguri, Noah Greenwald, Michael Angelo, Travis J Hollmann, and Saad Nadeem. 2021. DeepLIIF: Deep Learning-Inferred Multiplex ImmunoFluorescence for IHC Image Quantification. *bioRxiv* (2021).

[33] Fatemeh Ghezloo, Pin-Chieh Wang, Kathleen F Kerr, Tad T Brunyé, Trafton Drew, Oliver H Chang, Lisa M Reisch, Linda G Shapiro, and Joann G Elmore. 2022. An Analysis of Pathologists' Viewing Processes as They Diagnose Whole Slide Digital Images. *Journal of Pathology Informatics* (2022), 100104.

[34] Hongyan Gu, Mohammad Haeri, Shuo Ni, Christopher Kazu Williams, Neda Zarrin-Khameh, Shino Magaki, and Xiang'Anthony' Chen. 2022. Detecting Mitoses with a Convolutional Neural Network for MIDOG 2022 Challenge. *arXiv preprint arXiv:2208.12437* (2022).

[35] Hongyan Gu, Jingbin Huang, Lauren Hung, and Xiang 'Anthony' Chen. 2021. Lessons Learned from Designing an AI-Enabled Diagnosis Tool for Pathologists. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1, Article 10 (apr 2021), 25 pages. https://doi.org/10.1145/3449084

[36] Chu Han, Jiatai Lin, Jinhai Mai, Yi Wang, Qingling Zhang, Bingchao Zhao, Xin Chen, Xipeng Pan, Zhenwei Shi, Xiaowei Xu, et al. 2021. Multi-Layer Pseudo-Supervision for Histopathology Tissue Semantic Segmentation using Patch-level Classification Labels. *arXiv preprint arXiv:2110.08048* (2021).

[37] Sandra G. Hart and Lowell E. Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In *Human Mental Workload*, Peter A. Hancock and Najmedin Meshkati (Eds.). Advances in Psychology, Vol. 52. North-Holland, 139–183. https://doi.org/10.1016/S0166-4115(08)62386-9

[38] Narayan Hegde, Jason D Hipp, Yun Liu, Michael Emmert-Buck, Emily Reif, Daniel Smilkov, Michael Terry, Carrie J Cai, Mahul B Amin, Craig H Mermel, et al. 2019. Similar image search for histopathology: SMILY. *NPJ digital medicine* 2, 1 (2019), 1–9.

[39] Tad Hirsch, Kritzia Merced, Shrikanth Narayanan, Zac E. Imel, and David C. Atkins. 2017. Designing Contestability: Interaction Design, Machine Learning, and Mental Health. In *Proceedings of the 2017 Conference on Designing Interactive Systems* (Edinburgh, United Kingdom) *(DIS '17)*. Association for Computing Machinery, New York, NY, USA, 95–99. https://doi.org/10.1145/3064663.3064703

[40] Andreas Holzinger, André Carrington, and Heimo Müller. 2020. Measuring the quality of explanations: the system causability scale (SCS). *KI-Künstliche Intelligenz* (2020), 1–6.

[41] Yongxiang Huang and Albert Chi-shing Chung. 2018. Improving high resolution histology image classification with deep spatial fusion network. In *Computational Pathology and Ophthalmic Medical Image Analysis*. Springer, 19–26.

[42] Peter A Humphrey. 2004. Gleason grading and prognostic factors in carcinoma of the prostate. *Modern pathology* 17, 3 (2004), 292–306.

[43] Maia Jacobs, Jeffrey He, Melanie F. Pradier, Barbara Lam, Andrew C. Ahn, Thomas H. McCoy, Roy H. Perlis, Finale Doshi-Velez, and Krzysztof Z. Gajos. 2021. Designing AI for Trust and Collaboration in Time-Constrained Medical Decisions: A Sociotechnical Lens. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 659, 14 pages. https://doi.org/10.1145/3411764.3445385

[44] Youngseung Jeon, Seungwan Jin, and Kyungsik Han. 2021. FANCY: human-centered, deep learning-based framework for fashion style analysis. In *Proceedings of the Web Conference 2021*. 2367–2378.

[45] Youngseung Jeon, Seungwan Jin, Patrick C Shih, and Kyungsik Han. 2021. FashionQ: an ai-driven creativity support tool for facilitating ideation in fashion design. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–18.

[46] Patrick W Jordan, Bruce Thomas, Ian Lyall McClelland, and Bernard Weerdmeester. 1996. *Usability evaluation in industry*. CRC Press.

[47] Sasikiran Kandula and Jeffrey Shaman. 2019. Reappraising the utility of Google flu trends. *PLoS computational biology* 15, 8 (2019), e1007258.

[48] Saif Khairat, David Marc, William Crosby, Ali Al Sanousi, et al. 2018. Reasons for physicians not adopting clinical decision support systems: critical analysis. *JMIR medical informatics* 6, 2 (2018), e8912.

[49] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. 2020. Concept Bottleneck Models. In *Proceedings of the 37th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 119)*, Hal Daumé III and Aarti Singh (Eds.). PMLR, 5338–5348. https://proceedings.mlr.press/v119/koh20a.html

[50] Robert Krueger, Johanna Beyer, Won-Dong Jang, Nam Wook Kim, Artem Sokolov, Peter K Sorger, and Hanspeter Pfister. 2019. Facetto: Combining unsupervised and supervised learning for hierarchical phenotype analysis in multi-channel image data. *IEEE transactions on visualization and computer graphics* 26, 1 (2019), 227–237.

[51] Bokyung Lee, Taeil Jin, Sung-Hee Lee, and Daniel Saakes. 2019. SmartManikin: Virtual Humans with Agency for Design Tools. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3290605.3300814

[52] Min Hun Lee, Daniel P Siewiorek, Asim Smailagic, Alexandre Bernardino, and Sergi Bermúdez i Badia. 2021. A human-ai collaborative approach for clinical decision making on rehabilitation assessment. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–14.

[53] Martin Lindvall, Claes Lundström, and Jonas Löwgren. 2021. *Rapid Assisted Visual Search: Supporting Digital Pathologists with Imperfect AI*. Association for Computing Machinery, New York, NY, USA, 504–513. https://doi.org/10.1145/3397481.3450681

[54] Geert Litjens, Peter Bandi, Babak Ehteshami Bejnordi, Oscar Geessink, Maschenka Balkenhol, Peter Bult, Altuna Halilovic, Meyke Hermsen, Rob van de Loo, Rob Vogels, et al. 2018. 1399 H&E-stained sentinel lymph node sections of breast cancer patients: the CAMELYON dataset. *GigaScience* 7, 6 (2018), giy065.

[55] Xingyu "Bruce" Liu, Ruolin Wang, Dingzeyu Li, Xiang Anthony Chen, and Amy Pavel. 2022. CrossA11y: Identifying Video Accessibility Issues via Cross-Modal Grounding. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology* (Bend, OR, USA) *(UIST '22)*. Association for Computing Machinery, New York, NY, USA, Article 43, 14 pages. https://doi.org/10.1145/3526113.3545703

[56] David N Louis, Hiroko Ohgaki, Otmar D Wiestler, Webster K Cavenee, Peter C Burger, Anne Jouvet, Bernd W Scheithauer, and Paul Kleihues. 2007. The 2007 WHO classification of tumours of the central nervous system. *Acta neuropathologica* 114, 2 (2007), 97–109.

[57] David N Louis, Arie Perry, Pieter Wesseling, Daniel J Brat, Ian A Cree, Dominique Figarella-Branger, Cynthia Hawkins, HK Ng, Stefan M Pfister, Guido Reifenberger, et al. 2021. The 2021 WHO classification of tumors of the central nervous system: a summary. *Neuro-oncology* 23, 8 (2021), 1231–1251.

[58] David N Louis, Arie Perry, Pieter Wesseling, Daniel J Brat, Ian A Cree, Dominique Figarella-Branger, Cynthia Hawkins, H K Ng, Stefan M Pfister, Guido Reifenberger, Riccardo Soffietti, Andreas von Deimling, and David W Ellison. 2021. The 2021 WHO Classification of Tumors of the Central Nervous System: a summary. *Neuro-Oncology* 23, 8 (06 2021), 1231–1251. https://doi.org/10.1093/neuonc/noab106

[59] William M Lydiatt, Snehal G Patel, Brian O'Sullivan, Margaret S Brandwein, John A Ridge, Jocelyn C Migliacci, Ashley M Loomis, and Jatin P Shah. 2017. Head and neck cancers—major changes in the American Joint Committee on cancer eighth edition cancer staging manual. *CA: a cancer journal for clinicians* 67, 2 (2017), 122–137.

[60] Gregory Maniatopoulos, Rob Procter, Sue Llewellyn, Gill Harvey, and Alan Boyd. 2015. Moving beyond local practice: reconfiguring the adoption of a breast cancer diagnostic technology. *Social Science & Medicine* 131 (2015), 98–106.

[61] Melissa D McCradden, Shalmali Joshi, James A Anderson, Mjaye Mazwi, Anna Goldenberg, and Randi Zlotnik Shaul. 2020. Patient safety and quality improvement: Ethical principles for a regulatory approach to bias in healthcare machine learning. *Journal of the American Medical Informatics Association* 27, 12 (2020), 2024–2027.

[62] B Middleton, DF Sittig, and A Wright. 2016. Clinical decision support: a 25 year retrospective and a 25 year vision. *Yearbook of medical informatics* Suppl 1 (2016), S103.

[63] Jesper Molin, Morten Fjeld, Claudia Mello-Thoms, and Claes Lundström. 2015. Slide navigation patterns among pathologists with long experience of digital review. *Histopathology* 67, 2 (2015), 185–192.

[64] Lucio Palma, Paolo Celli, Carmine Franco, Luigi Cervoni, and Giampaolo Cantore. 1997. Long-term prognosis for atypical and malignant meningiomas: a study of 71 surgical cases. *Journal of neurosurgery* 86, 5 (1997), 793–800.

[65] Liron Pantanowitz, Paul N Valenstein, Andrew J Evans, Keith J Kaplan, John D Pfeifer, David C Wilbur, Laura C Collins, and Terence J Colgan. 2011. Review of the current state of whole slide imaging in pathology. *Journal of pathology informatics* 2 (2011).

[66] Sun Young Park, Pei-Yi Kuo, Andrea Barbarin, Elizabeth Kaziunas, Astrid Chow, Karandeep Singh, Lauren Wilcox, and Walter S Lasecki. 2019. Identifying challenges and opportunities in human-AI collaboration in healthcare. In *Conference Companion Publication of the 2019 on Computer Supported Cooperative Work and Social Computing*. 506–510.

[67] Alexander Rakhlin, Alexey Shvets, Vladimir Iglovikov, and Alexandr A Kalinin. 2018. Deep Convolutional Neural Networks for Breast Cancer Histology Image Analysis. In *Image Analysis and Recognition*. Springer International Publishing, Cham, 737–744.

[68] Ludovic Roux, Daniel Racoceanu, Nicolas Loménie, Maria Kulikova, Humayun Irshad, Jacques Klossa, Frédérique Capron, Catherine Genestie, Gilles Le Naour, and Metin N Gurcan. 2013. Mitosis detection in breast cancer histological images An ICPR 2012 contest. *Journal of pathology informatics* 4 (2013).

[69] Roy A Ruddle, Rhys G Thomas, Rebecca Randell, Philip Quirke, and Darren Treanor. 2016. The design and evaluation of interfaces for navigating gigapixel images in digital pathology. *ACM Transactions on Computer-Human Interaction (TOCHI)* 23, 1 (2016), 1–29.
[70] Cynthia Rudin. 2018. Please stop explaining black box models for high stakes decisions. *Stat* 1050 (2018), 26.
[71] Mike Schaekermann, Carrie J. Cai, Abigail E. Huang, and Rory Sayres. 2020. *Expert Discussions Improve Comprehension of Difficult Cases in Medical Image Assessment.* Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376290
[72] Mark Sendak, Madeleine Clare Elish, Michael Gao, Joseph Futoma, William Ratliff, Marshall Nichols, Armando Bedoya, Suresh Balu, and Cara O'Brien. 2020. "The Human Body is a Black Box": Supporting Clinical Decision-Making with Deep Learning. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 99–109. https://doi.org/10.1145/3351095.3372827
[73] David F Steiner, Kunal Nagpal, Rory Sayres, Davis J Foote, Benjamin D Wedin, Adam Pearce, Carrie J Cai, Samantha R Winter, Matthew Symonds, Liron Yatziv, et al. 2020. Evaluation of the Use of Combined Artificial Intelligence and Pathologist Assessment to Review and Grade Prostate Biopsies. *JAMA Network Open* 3, 11 (2020), e2023267–e2023267.
[74] Eliza Strickland. 2019. IBM Watson, heal thyself: How IBM overpromised and underdelivered on AI health care. *IEEE Spectrum* 56, 4 (2019), 24–31.
[75] Peter Ström, Kimmo Kartasalo, Henrik Olsson, Leslie Solorzano, Brett Delahunt, Daniel M Berney, David G Bostwick, Andrew J Evans, David J Grignon, Peter A Humphrey, et al. 2020. Artificial intelligence for diagnosis and grading of prostate cancer in biopsies: a population-based, diagnostic study. *The Lancet Oncology* 21, 2 (2020), 222–232.
[76] Randy L Teach and Edward H Shortliffe. 1981. An analysis of physician attitudes regarding computer-based clinical consultation systems. *Computers and Biomedical Research* 14, 6 (1981), 542–558.
[77] Hamid Reza Tizhoosh and Liron Pantanowitz. 2018. Artificial intelligence and digital pathology: challenges and opportunities. *Journal of pathology informatics* 9 (2018).
[78] Philipp Tschandl, Christoph Rinner, Zoe Apalla, Giuseppe Argenziano, Noel Codella, Allan Halpern, Monika Janda, Aimilios Lallas, Caterina Longo, Josep Malvehy, et al. 2020. Human–computer collaboration for skin cancer recognition. *Nature Medicine* 26, 8 (2020), 1229–1234.
[79] Mitko Veta, Yujing J Heng, Nikolas Stathonikos, Babak Ehteshami Bejnordi, Francisco Beca, Thomas Wollmann, Karl Rohr, Manan A Shah, Dayong Wang, Mikael Rousson, et al. 2019. Predicting breast tumor proliferation from whole-slide images: the TUPAC16 challenge. *Medical image analysis* 54 (2019), 111–121.
[80] Mitko Veta, Paul J Van Diest, Stefan M Willems, Haibo Wang, Anant Madabhushi, Angel Cruz-Roa, Fabio Gonzalez, Anders BL Larsen, Jacob S Vestergaard, Anders B Dahl, et al. 2015. Assessment of algorithms for mitosis detection in breast cancer histopathology images. *Medical image analysis* 20, 1 (2015), 237–248.
[81] Brian Patrick Walcott, Brian V Nahed, Priscilla K Brastianos, and Jay S Loeffler. 2013. Radiation treatment for WHO grade II and III meningiomas. *Frontiers in oncology* 3 (2013), 227.
[82] Dakuo Wang, Elizabeth Churchill, Pattie Maes, Xiangmin Fan, Ben Shneiderman, Yuanchun Shi, and Qianying Wang. 2020. From human-human collaboration to Human-AI collaboration: Designing AI systems that can work together with people. In *Extended abstracts of the 2020 CHI conference on human factors in computing systems*. 1–6.
[83] Dayong Wang, Aditya Khosla, Rishab Gargeya, Humayun Irshad, and Andrew H Beck. 2016. Deep learning for identifying metastatic breast cancer. *arXiv preprint arXiv:1606.05718* (2016).
[84] Dakuo Wang, Liuping Wang, Zhan Zhang, Ding Wang, Haiyi Zhu, Yvonne Gao, Xiangmin Fan, and Feng Tian. 2021. "Brilliant AI Doctor" in Rural Clinics: Challenges in AI-Powered Clinical Decision Support System Deployment. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–18.
[85] Shidan Wang, Donghan M Yang, Ruichen Rong, Xiaowei Zhan, Junya Fujimoto, Hongyu Liu, John Minna, Ignacio Ivan Wistuba, Yang Xie, and Guanghua Xiao. 2019. Artificial intelligence in lung cancer pathology image analysis. *Cancers* 11, 11 (2019), 1673.
[86] Nora S. Willett, Rubaiat Habib Kazi, Michael Chen, George Fitzmaurice, Adam Finkelstein, and Tovi Grossman. 2018. A Mixed-Initiative Interface for Animating Static Pictures. In *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology* (Berlin, Germany) *(UIST '18)*. Association for Computing Machinery, New York, NY, USA, 649–661. https://doi.org/10.1145/3242587.3242612
[87] Fuyong Xing, Toby C Cornish, Tell Bennett, Debashis Ghosh, and Lin Yang. 2019. Pixel-to-pixel learning with weak supervision for single-stage nucleus recognition in Ki67 images. *IEEE Transactions on Biomedical Engineering* 66, 11 (2019), 3088–3097.
[88] Zihan Yan, Yufei Wu, Yang Zhang, and Xiang 'Anthony' Chen. 2022. EmoGlass: An End-to-End AI-Enabled Wearable Platform for Enhancing Self-Awareness of Emotional Health. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 13, 19 pages. https://doi.org/10.1145/3491102.3501925
[89] Qian Yang, Aaron Steinfeld, Carolyn Rosé, and John Zimmerman. 2020. *Re-Examining Whether, Why, and How Human-AI Interaction Is Uniquely Difficult to Design.* Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376301
[90] Qian Yang, John Zimmerman, Aaron Steinfeld, Lisa Carey, and James F. Antaki. 2016. Investigating the Heart Pump Implant Decision Process: Opportunities for Decision Support Tools to Help. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing*

Systems (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 4477–4488. https://doi.org/10.1145/2858036.2858373

[91] Zizhao Zhang, Pingjun Chen, Mason McGough, Fuyong Xing, Chunbao Wang, Marilyn Bui, Yuanpu Xie, Manish Sapkota, Lei Cui, Jasreman Dhillon, et al. 2019. Pathologist-level interpretable whole-slide cancer diagnosis with deep learning. *Nature Machine Intelligence* 1, 5 (2019), 236–245.

A WHO GUIDELINES FOR MENINGIOMA GRADING

As specified by the World Health Organization (WHO) guidelines [56, 57], meningioma grading diagnosis can be based on the following criteria:

- **Grade 1** (benign) meningiomas include "histological variant other than clear cell, chordoid, papillary, and rhabdoid "[13] with some exceptions *and* a lack of criteria for grade 2 and 3 meningiomas.
- **Grade 2** (formerly called atypical) meningiomas are recognized by meeting at least one of the four following criteria:
  (1) The presence of $\geq$ 2.5 mitoses/mm^2 (equating to $\geq$ 4 mitoses per/10 high power field (HPF) of 0.16 mm^2. Moreover, since mitoses are challenging to recognize in H&E, the Ki-67-positive nuclei (Figure 3k) in the corresponding areas of Ki-67 (Figure 3d) are often compared for disambiguation;
  (2) At least three out of five following histopathological features are observed: hypercellularity — an abnormal excess of cells in the specimen (Figure 3f), prominent nucleoli — enlarged nucleoli in a cell (usually as a cluster) (Figure 3g,m), sheeting — loss of 'whirling' architecture (Figure 3h), necrosis — irreversible injury to cells (Figure 3i), and small cell — cluster of cells with high nuclear/cytoplasmic ratio (Figure 3j);
  (3) Brain invasion — invasive tumor cells within the brain tissue is observed (Figure 3e);
  (4) The dominant appearance of clear cell or chordoid subtype.
- **Grade 3** meningiomas are decided if at least one of the following criteria met [6, 57]:
  (1) Mitotic figures of $\geq$ 12.5 mitoses/mm ^{2} (equal to $\geq$ 20 mitoses/10 HPF of 0.16 mm^2);
  (2) The appearance of frank anaplasia, papillary or rhabdoid subtype with some exceptions;
  (3) Molecular alterations, such as a *TERT* promoter mutation; and/or homozygous *CDKN2A* and/or *CDKN2B* deletion.