Title

Supporting Diagnosis of Pathologists with Human-AI Collaboration

Author

Gu, Hongyan

Publication Date

2025

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Supporting Diagnosis of Pathologists with Human-AI Collaboration

A dissertation submitted in partial satisfaction

of the requirements for the degree

Doctor of Philosophy in Electrical and Computer Engineering

by

Hongyan Gu

2025

ABSTRACT OF THE DISSERTATION

Supporting Diagnosis of Pathologists with Human-AI Collaboration

by

Hongyan Gu
Doctor of Philosophy in Electrical and Computer Engineering
University of California, Los Angeles, 2025
Professor Xiang Chen, Chair

The recent trend of digital pathology transition has enabled the advancement of Artificial Intelligence (AI) for complex pathology tasks. While some AI demonstrated performance comparable to human pathologists in lab studies, translating them into clinical practice remains challenging due to issues related to limitations of AI's integration into clinical decision-making, its explainability and controllability, and the reliability of AI-assisted outcomes.

To address these challenges, this thesis adopts a multi-faceted approach, combining field investigations, artifact development, and empirical validation, to study effective human-AI collaborative paradigms in digital pathology. First, it presents findings from a field study of pathologists' daily workflows, their attitudes towards AI with varying levels of automation, and recommendations for designing effective AI-assisted diagnostic systems. Second, this thesis discusses the development and validation of NaviPath, a next-generation, high-throughput AI recommendation system informed by pathologists' domain expertise, and xPath, a comprehensive and explainable AI-assisted pathology interface that seamlessly integrates with pathologists' diagnostic tasks involving multiple criteria and multimodal data. Finally, this thesis explores strategies to foster appropriate reliance on AI by harnessing

pathologists' collective expertise to achieve reliable, and robust AI-assisted outcomes.

Overall, this thesis aspires to enable efficient, accurate, and safe human-AI collaborative pathology decisions – supporting pathologists in reaching timely, cost-effective, and precise diagnoses, which can ultimately benefit patient management.

The dissertation of Hongyan Gu is approved.

Lei He

Corey Wells Arnold

Lin Yang

Xiang Chen, Committee Chair

University of California, Los Angeles

2025

To my parents, Yali Han and Huiping Gu.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

Foremost, I would like to express my deepest gratitude to my advisor, Prof. Xiang 'Anthony' Chen, for his unconditional support, insightful guidance, and continuous encouragement throughout my Ph.D. journey. I am especially thankful to Prof. Mohammad Haeri from Kansas University Medical Center (KUMC) for introducing me to this magnificent field of pathology and providing invaluable perspectives that shaped my research direction and career development.

I am grateful to my thesis committee members, Prof. Lei He, Prof. Corey Arnold, and Prof. Lin Yang, for their valuable guidance and feedback on my research and this thesis.

None of the projects in this thesis would have been possible without the contribution of our outstanding pathologist collaborators: Dr. Harry V. Vinters, Dr. Shino Magaki, Dr. Negar Khanlou (UCLA Neuropathology); Dr. Neda Zarrin-Khameh (Baylor College of Medicine); Dr. Xinhai Robert Zhang (UTHealth Houston); Dr. Inma Cobos (Stanford); Dr. Carrie Mohila (Texas Children's Hospital); Dr. Ashley Holloman (Baylor College of Medicine); Dr. Joshua Byers (UCSF); Dr. Sara Stone (previously at University of Pennsylvania); Dr. Nelli Lakis, Dr. Jasmeet Assi, Dr. Ameer Hamza, Dr. Anders Meyer, Dr. Todd Stevens, Dr. Ethar Husseinawi, Dr. Hamilton Seth, and Dr. Amin Mostofizadeh (KUMC); Dr. Jing Wang (previously at CNSI), and Dr. Karam Han (University of Wisconsin).

I am also thankful to our dedicated pathologist trainee collaborators: Ellis McCormick, Ellie Onstott (KUMC); Dr. Hilda Mirbaha and Dr. Zesheng Chen (UCLA); Dr. Sallam Alrosan, Dr. Issa Al-Kharouf, Dr. Jwan Alallaf, Dr. Sarah Cain, Dr. Thuy Cao, Dr. Hebatullah Elsafy, Dr. Hanan Elsarraj, Dr. Sydney Graham, Dr. Julie Pham, and Dr. Janna Should (KUMC); and Dr. Mengxue Zhang (University of Chicago).

I would also like to thank my engineering collaborators: Prof. Dr. Marc Aubreville (previously at Technische Hochschule Ingolstadt, Germany), Dr. Yang Li (Google), Dr. Wenzhong Yan (Yale); Dr. Yuan Liang, Dr. Ruolin Wang, Tengyou Xu, Christian Giron-

Michel, Chunxu Yang, Zida Wu, Shuo Ni, Yifan Xu, Jingbin Huang, Lauren Hung (previously at UCLA ECE); Shirley Tang, Ayesha Alvi, Uyenvy Nguyen, Brandon Day (UCLA Cognitive Science).

Moreover, my Ph.D. journey would not have been half as enjoyable without my friends at UCLA ECE: Prof. Yang Zhang, Dr. Junbo Wang, Dr. Noyan Evirgen, Siyou Pei, Dr. Jiahao 'Nick' Li, Xiaoying Yang, and Xue Wang.

Most importantly, I would like to thank my parents, Yali Han and Huiping Gu, for their unwavering support, understanding, and encouragement throughout my academic journey. This thesis is dedicated to them.

Finally, I would like to express my sincere gratitude to all the anonymous patients who agreed to contribute to medical research. Their trust and participation make the advancement of medical science possible.

May Science and Rationality bring us closer to a future where cancer becomes a history.

VITA

2013–2017 B.Eng. in Automation, Zhejiang University.

2017–2019 M.S. in Electrical and Computer Engineering, University of California, Los Angeles.

2019–present Ph.D. student in Electrical and Computer Engineering, University of California, Los Angeles.

PUBLICATIONS

Gu, Hongyan, Jingbin Huang, Lauren Hung, and Xiang 'Anthony' Chen. "Lessons learned from designing an AI-enabled diagnosis tool for pathologists." Proceedings of the *ACM on Human-computer Interaction* 5, no. CSCW1 (2021): 1-25.

Gu, Hongyan, Mohammad Haeri, Shuo Ni, Christopher Kazu Williams, Neda Zarrin-Khameh, Shino Magaki, and Xiang 'Anthony' Chen. "Detecting mitoses with a convolutional neural network for midog 2022 challenge." In MICCAI Challenge on Mitosis Domain Generalization, pp. 211-216. Cham: Springer Nature Switzerland, 2022.

Gu, Hongyan, Yuan Liang, Yifan Xu, Christopher Kazu Williams, Shino Magaki, Negar Khanlou, Harry Vinters, Yang Li, Mohammad Haeri, and Xiang 'Anthony' Chen "Improving workflow integration with XPath: design and evaluation of a human-AI diagnosis system in pathology." *ACM Transactions on Computer-Human Interaction* 30, no. 2 (2023): 1-37.

Gu, Hongyan, Chunxu Yang, Mohammad Haeri, Jing Wang, Shirley Tang, Wenzhong Yan, Shujin He, Christopher Kazu Williams, Shino Magaki, and Xiang 'Anthony' Chen. "Augmenting pathologists with NaviPath: design and evaluation of a human-AI collaborative navigation system." In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pp. 1-19. 2023.

Gu, Hongyan, Chunxu Yang, Issa Al-Kharouf, Shino Magaki, Nelli Lakis, Christopher Kazu Williams, Sallam Mohammad Alrosan, Ellie Kate Onstott, Wenzhong Yan, Negar Khanlou, Inma Cobos, Xinhai Robert Zhang, Neda Zarrin-Khameh, Harry V Vinters, Xiang 'Anthony' Chen, and Mohammad Haeri. "Enhancing mitosis quantification and detection in meningiomas with computational digital pathology." *Acta Neuropathologica Communications* 12, no. 1 (2024): 7.

Gu, Hongyan, Zihan Yan, Ayesha Alvi, Brandon Day, Chunxu Yang, Zida Wu, Shino Magaki, Mohammad Haeri, and Xiang 'Anthony' Chen. "Supporting Mitosis Detection AI Training with Inter-Observer Eye-Gaze Consistencies." In *2024 IEEE 12th International Conference on Healthcare Informatics (ICHI)*, pp. 40-45. IEEE, 2024.

Gu, Hongyan, Chunxu Yang, Shino Magaki, Neda Zarrin-Khameh, Nelli S. Lakis, Inma Cobos, Negar Khanlou, Xinhai R Zhang, Jasmeet Assi, Joshua T Byers, Ameer Hamza, Karam Han, Anders Meyer, Hilda Mirbaha, Carrie A Mohila, Todd M Stevens, Sara L Stone, Wenzhong Yan, Mohammad Haeri, and Xiang 'Anthony' Chen "Majority voting of doctors improves appropriateness of AI reliance in pathology." *International Journal of Human-Computer Studies* (2024): 103315.

CHAPTER 1

Introduction

1.1 Background and Motivation

Driven by hospitals' urging demands for transitioning to digital workflows during the COVID-19 pandemic, digital pathology has been broadly applied in research and industry over the past five years [97, 150, 225]. This surge in data availability has stimulated the advancement of self-supervised training methods for pathology foundation models. By 2025, a popular approach in pathology Artificial Intelligence (AI) is through unsupervised pre-training on mass datasets (*e.g.*, Virchow, with 1.5 million slides [207]) followed by fine-tuning on downstream tasks. This methodology has enabled deep learning models to accomplish complex pathology tasks with performance non-inferior to human pathologists, including classifying tumor subtypes, grades, predicting immunohistochemistry (IHC), molecular mutation, and prognosis [3, 54, 214, 207, 218].

Despite these advances, the clinical translation of digital pathology AI faces several critical hurdles [89, 44, 169, 220]. First, the misalignment between data scientists' AI design and pathologists' workflow dynamics raises concerns about the actual clinical utility of AI: can these AI systems truly augment pathologists' workflow, or do they inadvertently introduce new cognitive burdens? Second, existing deep learning approaches are often criticized for their lack of interpretability and transparency, making their behavior unpredictable and fostering skepticism among pathologists. Finally, despite the high-stakes nature of pathology decision-making, quality control of AI-assisted results relies majorly on human pathologists

as the gatekeepers: how can we establish rigorous mechanisms for quality assurance that mitigate the risks of AI-induced errors while preserving pathologists' efficiency?

These sociotechnical challenges at the intersection of AI and clinical practice present a unique opportunity to fundamentally rethink the design of AI ecosystems for supporting pathologists' diagnoses. Rather than applying AI as a stand-alone tool, there is a crucial need to explore *human-AI collaborative paradigms* that seamlessly integrate into pathologists' diagnostic workflows while preserving their expertise and decision-making autonomy. This thesis fills this gap by systematically analyzing pathologists' examination behaviors, understanding their expectations and preferences for AI integration, developing explainable, integrable, and customizable AI assistance that incorporates pathologists' domain knowledge, and harnessing the collective intelligence of AI-assisted pathologists to mitigate AI-induced errors. In doing so, *this thesis aims to establish a new paradigm in digital pathology – one that fosters efficient, accurate, and safe human-AI collaboration, enabling timely, cost-effective, and precise histopathological diagnoses that ultimately enhance clinical decision-making and improve patient management.*

1.1.1 Thesis Statement

AI-enabled digital pathology interfaces can improve pathologists' speed and decision outcome while mitigating AI-induced errors through three key innovations: *(i)* a next-generation, high-throughput region-of-interest AI recommendation system for a challenging navigation task, informed by pathologists' domain expertise, *(ii)* a comprehensive and explainable AI digital pathology interface that aligns with pathologists' diagnostic workflows for a complex diagnosis task, and *(iii)* a majority voting strategy that ensembles collective, AI-assisted pathologists' decisions. These innovations are validated through comprehensive user study sessions with pathologists and trainees, demonstrating improvements in examination efficiency, inter-observer agreement, decision accuracy, appropriateness, and robustness, compared to traditional manual and AI alone.

1.2 Related Work

This section provides a brief review of recent literature in four directions related to this thesis: *(i)* digital pathology technologies supporting pathologists' visual examination, *(ii)* development and application of pathology AI, *(iii)* human-AI collaborative systems for pathology applications, and *(iv)* enabling appropriate AI reliance for high-stakes pathology applications.

1.2.1 Supporting Pathologists' Examination with Digital Pathology

Since the U.S. Food and Drug Administration (FDA) granted the first 510(k) clearance for a digital pathology scanning system in 2017 [73], a wide variety of digital pathology solutions have emerged and achieved commercial implementation [97, 150, 225]. Starting from the late 2010s, numerous medical centers have successfully transitioned to full-digital workflows, bringing benefits including the convenience of remote sign-out, consultation, and education, reduced time in case assembly and delivery, higher efficiency in slide reading, and lower injury rates related to slide handling [38, 71, 98, 170, 176]. As of 2025, digital pathology technology encompasses the majority of formalin-fixed paraffin-embedded (FFPE) slides, with exceptions in frozen sections, cytology, and non-FFPE hematopathology. A typical digital pathology system consists of three core components: *(i)* the digital pathology scanner [160], *(ii)* the pathology Image Management System (IMS), and *(iii)* displays [1]. As the essential visualization component between pathologists and digital slides, the display and interface design play a pivotal role in pathologists' efficiency during histological examination, which directly impacts their digital experience and diagnostic outcome.

Nowadays (2025), commercial off-the-shelf displays can have up to 8K resolution (7680 $\times$ 4320, or $\sim 3.3 \times 10^7$ pixels), which still remains significantly lower than pathology scans (up to 10^{12} pixels). Therefore, intensive navigation is usually required for pathologists to search for small-sized histopathological patterns (*e.g.*, mitoses, Helicobacter pylori) during

examination [174]. As this limitation is majorly due to the resolution differences, one intuitive solution is to introduce displays with larger physical sizes and resolutions to pathologists [165, 164, 197, 215]. Literature has validated this approach, demonstrating that pathologists had less pan and zoom interactions with higher-resolution displays [142]. However, improving hardware requires purchasing costly, bulky, and specialized devices. As such, I believe that an optimization of interface design presents a more practical approach to to support pathologists in working efficiently with high-resolution digital scans.

In the human-computer interaction domain, extensive studies have investigated design strategies to support users in navigating high-resolution images with limited size screens or displays [25, 94, 177, 171, 224]. Cockburn *et al.*categorized these strategies into four fundamental approaches: focus + context (F+C), overview + detail (O+D), zooming, and cue-based [57]. In digital pathology, numerous open-source [20, 59, 155, 181] and commercial [110] interfaces adapt a design that combines zooming and O+D, featuring a zoomable canvas showing pathology scan details and an overview window displaying the thumbnail. Users navigate high-resolution images with "pan and zoom" [78] interactions. However, criticisms suggest that such design demands a high mental effort and might be time-consuming [115, 174]. To address this limitation, Randell *et al.* improved the design by enlarging the overview to detail scale difference, enabling pathologists to pan more efficiently by moving the cursor in the 'overview' window [174]. Beyond O+D approaches, Jessup *et al.* proposed an F+C interface for pathology image exploration [115]: a focal lens that magnifies the screen center and supports users' close-up examinations and explorations of multi-channeled pathology scans.

1.2.2 AI Technologies for Pathology Applications

With the recent application of digital pathology, massive amounts of pathology image data has been accumulated. Large-scale, open-source unlabeled datasets, such as E-brains [172]

and the NIH Cancer Genome Atlas (TCGA)[1], along with substantial volumes of digital slides generated by early-adopting centers (PB-level/center/year [97]), have laid the foundation for pathology AI, or computational pathology research. Furthermore, a considerable number of annotated datasets have emerged, covering a broad range of pathology specialties, from conducting high-level diagnostic tasks, such as identifying breast cancer metastasis [132], detecting perineural invasion [55], grading prostate cancer [41], to recognizing low-level histological patterns, such as mitoses [12, 13, 14, 204, 173, 194], cell phenotype [5, 86], and necrosis [6].

The increasing availability of digital pathology data has triggered a surge in data-driven techniques in the past decade, which revolutionizes broad aspects of digital pathology research [185]. The reliability and diagnostic accuracy of these approaches have been systematically evaluated in recent reviews [144]. These techniques have demonstrated promising results across a wide range of pathology applications, including screening negative specimens [66, 35], carcinoma detection [8, 27, 23], quantification of histological features [56, 206], immunohistochemistry (IHC) analysis [82], tumor grading [69, 10, 167, 213], predicting molecular changes [153], and prognosis prediction [223, 214].

Notably, an emerging and popular approach for computational pathology is to train foundation models in an unsupervised manner using millions of digital pathology images, followed by fine-tuning these models for downstream tasks, as demonstrated by UNI [54], CHIEF [214], Gigapath [218] and Virchow [207]. Meanwhile, with the advancement of Large Language Models (LLMs), researchers have begun exploring vision-language models trained on publicly available data sources, such as social media posts [107] and scientific literature [138], establishing a novel direction for training "pathology AI companion" with reasonable cost. Despite this promising trend, the reliance on private training data in computational pathology poses challenges for institutions behind digitization progress to develop their own AI solutions. However, recent breakthroughs in efficient training methodologies, as demon-

[1] `https://www.cancer.gov/ccg/research/genome-sequencing/tcga`

strated by DeepSeek [93], may offer new perspectives for cost-effective pathology foundation model development in the future.

Furthermore, due to legislative and ethical concerns, as of 2025, AI algorithms cannot replace pathologists' examinations, neither they may work as a "stand-alone" application in clinical practice [53, 203]. Instead, they are regarded as "software as a medical device" to assist doctors[2], with one designed for prostate cancer pathology receiving the first FDA *de novo* approval in 2021 [74].

1.2.3 Human-AI Collaboration for Medical Decision-Making

The concept of human-AI collaboration envisions human-machine symbiosis [130], where humans and machines work together to achieve mutual goals [209]. Recent advances in deep learning have established foundational research in human-AI collaboration, suggesting the principles [105], guidelines [4], design recommendations [89], and information needs [44] for effective collaborative paradigms.

Building upon these foundations, researchers have explored human-AI collaboration for medical decision-making. For example, Beede *et al.* discovered socio-environmental factors that can influence AI performance, nurses' workflows, and patient experiences while deploying a deep learning model to detect diabetic retinopathy [26]. Wang *et al.* concluded the challenges of applying a clinical diagnostic support system in rural clinics [210]. Lee *et al.* proposed a human-AI collaboration system for therapists' practices of rehabilitation assessments, and reported that the system can increase the consistency of decision-making [127]. More recently, Fogliato *et al.* have studied the influence of human-AI workflows on veterinary radiologist readings of X-ray images, and revealed that doctors' findings were more aligned if AI suggestions were shown from the beginning [76]. Schaekermann *et al.* discovered that implementing ambiguity-aware AI was more effective in guiding medical experts' at-

[2] `https://www.fda.gov/media/145022/download`

tention to contentious portions while reviewing sheep EEG data, compared to conventional AI [178]. Calisto *et al.* extended the designs of multi-modality radiology image viewing tools [45, 46]. They built clinician-AI workflows for breast cancer image classification, suggesting that the human + AI approach could bring improvements in false-positives and false-negatives in diagnosis, user satisfaction, and time consumption [47, 48].

In the pathology domain, human-AI collaborative approaches have shown improvements in error reduction [211, 74, 108], between-subject agreements [19, 40, 200], time consumption [133, 113, 201, 16], mental workload and confidence [90, 92]. Lindvall *et al.* adapted the notion of Rapid Serial Visual Presentation [186] and developed a rapid, AI-assisted visual search system, allowing pathologists to see and adjust the AI-generated ROIs by sensitivity [113]. Cai *et al.* built a pathology CBIR system with an imperfect AI model – pathologists could adjust the retrieved ROIs according to pathologist-defined concepts (*e.g.*, stroma) to cope with AI imperfections [43]. More recently, Huang *et al.* proposed a human-in-the-loop framework that enables pathologists to train a machine learning model interactively by providing annotations, which achieved significantly higher identification performance in detecting plasma cells, colorectal carcinoma, and lymph node metastasis [108].

A crucial aspect of human-AI collaboration in pathology is achieving complementary team performance, where the collaboration outperforms both individual pathologists and AI systems [22].: In 2016, Wang *et al.* theoretically demonstrated this possibility by showing reduced error rates in breast cancer classification by combining AI and pathologist decisions [211]. However, there is a lack of empirical evidence to support it in pathology. Moreover, several studies in the general domains have failed to observe the task accuracy improvement in human-AI collaboration compared to AI alone [22, 125, 114]. This issue may be due to users' misuse of AI, a factor that can negatively impact the outcome of human-AI collaboration [120, 49, 202].

1.2.4 Enabling Appropriate AI Reliance

According to [159, 202, 180], two goals should be achieved to enable appropriate AI reliance: *(i)* mitigating over-reliance, where humans can identify and reject AI's incorrect recommendations, and *(ii)* reducing under-reliance, where humans can overcome their aversion of AI and accept its correct recommendations.

For enabling appropriate AI reliance, this is a tendency in research to study mechanisms and counter-measures for over-reliance. For instance, the cognitive forcing function, which prompts users to think analytically before decision-making, has shown promise [39]. Similarly, altering the interaction speed, where enlonging the AI response time, can instigate users' reflective thinking. Therefore, over-reliance incidents could be reduced [158, 166, 126]. Other approaches aim to enhance users' onboarding process, such as improving AI literacy [124, 135, 128], where users are informed of AI details [44, 113, 114]. However, translating these approaches to the medical domain may encounter two challenges. Firstly, introducing cognitive forcing functions or altering interaction speed could develop 'algorithm aversion,' especially when medical tasks are time-sensitive [68, 76]. Secondly, the efficacy of enhancing AI literacy also appeared marginal, possibly because of the difficulties in educating users within a limited timeframe [124, 128].

Besides these, another popular approach is eXplainable AI (XAI), where over-reliance might be reduced by enabling users to understand AI's reasoning [42, 124, 227, 22]. Nonetheless, numerous studies have failed to observe this anticipated effectiveness [179]: The potential benefits of XAI may be offset by the cognitive efforts of interpreting them [202]. Given the already high cognitive demands of medical professionals, this might result in XAI being less referred to, countering its potential benefits. This issue of appropriateness usage of XAI in medicine was raised by [104]. Further research suggested causability, an ability of an explanation that can enable casual understanding of medical experts, should also considered and measured to achieve better efficiency, effectiveness, and user satisfaction [103, 161].

1.3 Challenges and Research Questions

Despite the promising progress in pathology AI research, their translation to clinical implementation, particularly in developing AI systems that gain pathologists' acceptance and provide practical utility, presents significant challenges. These challenges are primarily due to the intrinsic characteristics of modern (*i.e.*, post 2010s) AI, whose limitations manifest in three critical aspects: *(i)* poor controllability with the unpredictable outcome; *(ii)* the inherent opacity of their inferencing processes (commonly referred to as the "black-box" problem), and *(iii)* uncertain generalizability across various scanners and stains [169]. Moreover, as previously discussed, ethical considerations pose barriers to AI deployment as stand-alone tools in clinical settings.

Researchers have long realized the limitation of using AI as a "Greek Oracle", and pointed out "*physician-user and the consultant program should interact symbiotically*" [147]. Although various works have concluded the suggestions and guidelines [4, 105] on how humans and AI should work collaboratively in the Human Computer Interaction (HCI) community, previous work suggests that it is "uniquely difficult" [220] to translate AI-augmented systems to clinical applications. The HCI problem of pathology using AI is its poor workflow integration due to the large knowledge gap between the two domains: pathology is highly specialized domain in medicine, requiring specific expert domain knowledge and strategies [149, 174] to assist doctors' decisions. As state-of-the-art AI focuses on pushing the performance with data-driven, 'end-to-end' models, pathologists' need for an AI's workflow integration is more or less ignored, which disincentives them from accepting and using AI in practice [222]. Even if pathologists develop trust in AI, another critical challenge arises: how to foster a safe and reliable environment that enables pathologists to develop appropriate AI reliance – that is, accepting correct AI recommendations while rejecting incorrect ones. Existing approaches to regularizing AI reliance – such as XAI and "cognitive forcing functions" – inevitably disrupt pathologists' natural examination behavior, which limits their utility in real-world clinical

settings.

To address these challenges, this thesis explores human-AI collaboration paradigms for support pathologists' diagnoses – from both qualitative and quantitative perspectives – aiming to bridge the knowledge gap between AI and clinical pathology practice. Specifically, this thesis seeks to answer the following research questions:

RQ1 How should human-AI collaboration systems be designed for pathology, and how can these insights inform future system development?

RQ2 How does human-AI collaboration affect pathologists' examination and diagnostic processes?

RQ3 How can human-AI collaboration be optimized to maximize pathologists' correctness while ensuring appropriate AI reliance?

1.4 Outline

The remainder of this thesis is organized as follows:

- **Chapter 2** presents qualitative findings from a field study investigating pathologists' attitudes toward AI automation in the context of fine-tuning a non-perfect model for lymph node metastasis detection. This chapter introduces Impetus, a proof-of-concept prototype that enables pathologists to interactively refine AI by providing coarse annotations through mixed-initiative designs. Based on pathologists' feedback, this chapter synthesizes six recommendations for designing human-AI collaborative systems to align with pathologists' real-world needs.

- **Chapter 3** designs and evaluates a human-AI collaborative workflow informed by pathologists' navigation domain knowledge. It presents NaviPath, a high-throughput

AI-assisted navigation system that provides explainable and customizable region-of-interest AI recommendations. A user study with 15 pathologists and residents was conducted to test the efficacy of NaviPath in challenging navigation tasks, and NaviPath's performance was compared to both manual examination and AI alone.

- **Chapter 4** identifies three key AI challenges in complex pathology decision-making processes: *comprehensiveness*, *explainability*, and *integrability*. To overcome these, this chapter manifests a top-down human-AI collaboration paradigm by developing xPath, a meningioma grading tool that supports the examination of multiple criteria and multimodal Hematoxylin and Eosin – IHC slides. A user experience study with 12 pathologists and residents evaluates xPath's usability and its impact on diagnostic outcomes.

- **Chapter 5** investigates a strategy for fostering appropriate AI reliance in high-stakes pathology decision-making. It examines how pathologists interact with AI and XAI assistance in mitosis detection and evaluates the impact of a majority voting mechanism that ensembles multiple AI-assisted pathologists' decisions. A nationwide user study with 32 pathologists and trainees provides empirical results to reveal whether and why majority voting improves conventional AI-assisted decisions.

- **Chapter 6** concludes this thesis by discussing key findings and potential directions for further advancing human-AI collaboration in digital pathology.

CHAPTER 2

Understanding Human-AI Collaborative Workflows in Pathology

This chapter is based in part on the following publication:

Hongyan Gu, Jingbin Huang, Lauren Hung, and Xiang 'Anthony' Chen. "Lessons learned from designing an AI-enabled diagnosis tool for pathologists." Proceedings of the *ACM on Human-computer Interaction* 5, no. CSCW1 (2021): 1-25.

2.1 Introduction

In this chapter, we propose a series of physician-AI collaboration techniques, based on which we prototype Impetus — a tool where an AI aids a pathologist in histological slide tumor detection using multiple degrees of initiative. Trained on a limited-sized dataset, our AI model cannot fully automate the examination process; instead, Impetus harnesses AI to *(i)* guide pathologists' attention to regions of major outliers, thus helping them prioritize the manual examination process; *(ii)* use agile labeling to train and adapt itself on-the-fly by learning from pathologists; and *(iii)* take initiatives appropriately for the level of performance confidence, from full automation, to pre-filling diagnosis, and to defaulting back to manual examination. We used the Impetus prototype as a medium to engage pathologists and observe how they perform diagnosis with AI involved in the process and elicit pathologists' qualitative reactions and feedback on the aforementioned collaborative techniques. From work sessions with eight pathologists from a local medical center, we summarize lessons

learned as follows.

Lesson #1 To explain AI's guidance, suggestions and recommendations, the system should go beyond a one-size-fits-all concept and provide instance-specific details that allow a medical user to see evidence that leads to a recommendation.

Lesson #2 Medical diagnosis is seldom a one-shot task, thus AI's recommendations need to continuously direct a medical user to filter and prioritize a large task space, taking into account new information extracted from a user's up-to-date input.

Lesson #3 Medical tasks are often time-critical, thus the benefits of AI's guidance, suggestions and recommendations need to be weighed by the amount of extra efforts incurred and the actionability of the provided information.

Lesson #4 To guide the examination process with prioritization, AI should help a medical user narrow in small regions of a large task space, as well as helping them filter out information within specific regions.

Lesson #5 It is possible for medical users to provide labels during their workflow with acceptable extra effort. However, the system should provide explicit feedback on how the model improves as a result, as a way to motivate and guide medical users' future inputs.

Lesson #6 Tasks treated equally by an AI might carry different weights to a medical user. Thus for medically high-staked tasks, AI should provide information to validate its confidence level.

Importantly, these lessons reveal what was unexpected as pathologists collaborated with AI using Impetus' techniques, which we further discuss as design recommendations for the future development of human-centered AI for medical imaging.

2.1.1 Contributions

Our contributions are as follows.

- The first suite of interaction techniques in medical diagnosis that instantiate mixed-initiative principles [105] for physicians to interact with AI with adaptive degree of initiatives based on AI's capabilities and limitations;
- A proof-of-concept system that embodies these techniques as an integrated diagnostic tool for pathologists to detect tumors from histological slides;
- A summary of observations and lessons learned from a study with eight pathologists that provides empirical evidence of employing mixed-initiative interaction in the medical imaging domain, thus informing future work on the design and development of human-centered AI systems.

2.2 Medical Background

In our study, we used a dataset containing Hematoxylin and Eosin (H&E) stained sentinel lymph node (SLN) sections of breast cancer patients [132]. The diagnosis of such specimens contains four main categories [7]:

- **Isolated tumor cells (ITC)** if the node contains a single tumor cell or cell deposits that are no larger than 0.2 mm or contain fewer than 200 cells;
- **Micro** if containing metastasis greater than 0.2 mm or more than 200 cells;
- **Macro** if containing metastasis greater than 2 mm;
- **Negative** if containing no tumor cells.

2.3 Scenario Walkthrough of Impetus

The user of Impetus, a pathologist, starts diagnosing a patent's case by importing multiple Whole Slide Images (WSI) of the patient into Impetus.

First, the pathologist's attention is drawn to the two boxes generated by the AI, which encompass regions of patches that visually appear to be 'outliers' from the majority of cells (Figure 2.1(a)), which suggests that these patches are likely to be tumor-positive. With these automatic recommendations, Impetus alleviates the pathologist's burden of navigating a large, high-resolution image and having to go through a large number of areas that might or might not be as tumor-characteristic as the recommended regions.

Next, the pathologist performs diagnosis by marking each recommended region as either 'tumor' or 'normal', and continues to marquee-select and label a few more regions on the WSI (Figure 2.1(b)). As the pathologist makes such selections, their input is also collected by the back end AI and used as labels to adapt the model better to align itself with the pathologist's domain knowledge.

Based on these diagnostic inputs and revisions from the pathologist, Impetus immediately adapts the underlying AI model accordingly. In contrast to conventional data labeling tasks, Impetus's agile labeling is designed to be lightweight and can learn from pathologists' input of coarsely marked regions without having to trace a precise contour of a tumor region. In this way, Impetus allows pathologists to agilely train an AI model as a natural and integral part of their existing workflow without incurring extra effort.

As the pathologist annotates more WSIs (which also trains the AI), they notice that some new slides are already marked as 'diagnosed' — AI takes the initiative to diagnose slides that it feels highly confident about. Thus the pathologist skips ahead to see other unlabeled slides, some of which, have pre-filled diagnosis dialogues (Figure 2.1(c)). In such cases, the pathologist examines the WSI to verify the AI's hypothesis. In the rest of the WSIs, the AI almost becomes invisible (due to a lack of confidence), and the pathologist

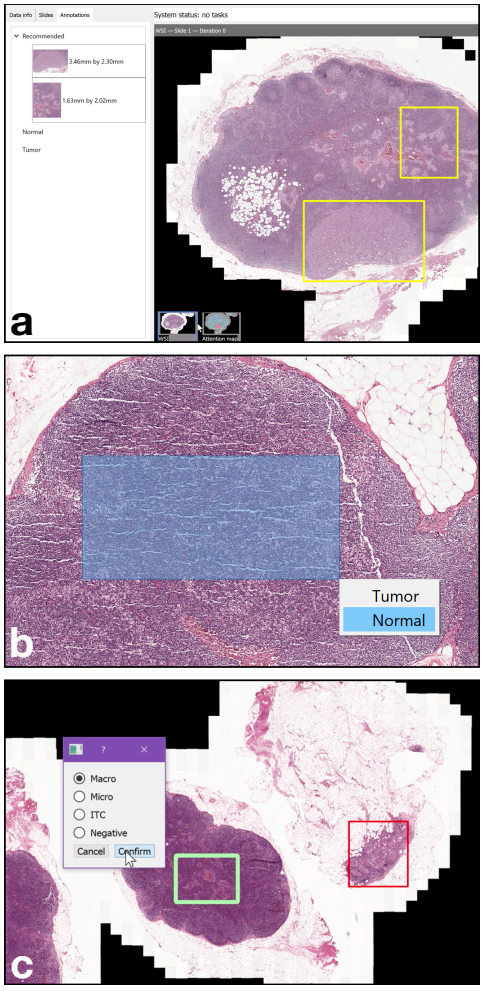


Figure 2.1: Key interactive features of IMPETUS: (a) as a pathologist loads a whole slide image, AI highlights areas of interest identified by outlier detection, shown as two yellow recommended boxes. (b) Agile labeling: a pathologist can drag and click to provide a label that can be employed to train the AI's model. (c) Diagnosis dialogue, pre-filled with AI's diagnosis, allows the pathologist to either confirm or disregard and proceed with manual diagnosis.

proceeds to to finish the diagnostic tasks manually.

The above scenario demonstrates how an 'imperfect' AI can still benefit a pathologist without necessarily automating the user's existing workflow: recommendation boxes suggestively prioritize pathologists' manual searching process (Figure 2.1(a)), agile labeling adapts AI while minimizing the extra effort from the pathologists (Figure 2.1(b)), and as AI attempts to improve itself, it handles cases with different degrees of initiatives — from full automation to pre-filling plausible results to remaining complete 'invisible'—based on its confidence (Figure 2.1(c)).

2.4 Design and Implementation

Below we first describe the design process then detail the specific interaction techniques and their implementation in Impetus.

2.4.1 Overview of the Design Process: Empirical and Theoretical Grounding

The design of Impetus is grounded in both empirical evidence and principles drawn from literature.

On the empirical side, we co-designed Impetus with our pathologist collaborator. Specifically, we learned that one major challenge for pathologists is efficiently and effectively navigating large, high-resolution WSIs. This suggests that AI, besides making diagnosis, can usefully serve to guide pathologists to navigate complex and high-resolution image space. We detail this design in Chapter 2.4.2.

On the theoretical side, Impetus goes beyond the singular objective of automation by offering a spectrum of AI-enabled assistance. As pointed out by Blois' seminal paper [33], a physician's differential diagnosis process is similar to a funnel, starting with a broad exploration of plausible conditions and gradually rule out less likely possibilities as more

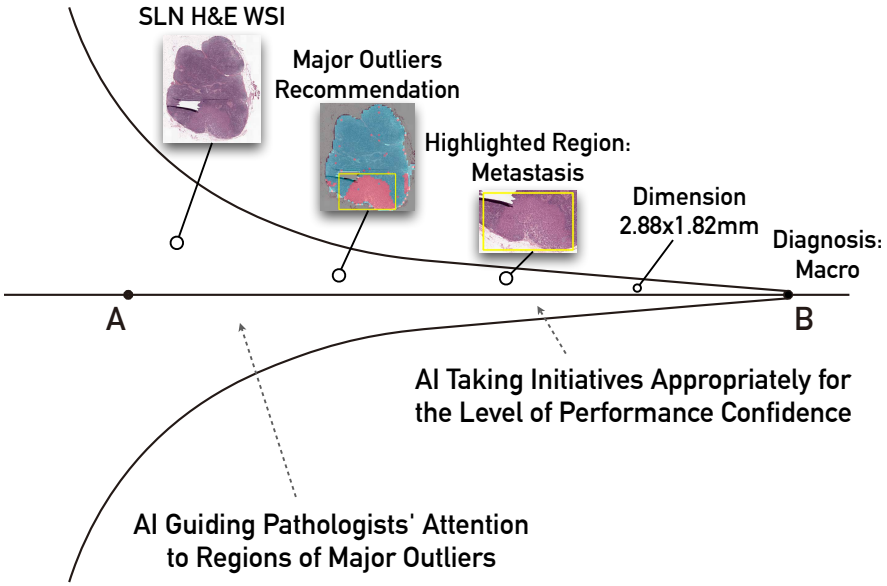


Figure 2.2: A physician's differential diagnosis process is similar to a funnel, starting with a broad exploration of plausible conditions and gradually rule out less likely possibilities as more evidence (*e.g.*, test results) is gathered until finally a single most probable conclusion can be drawn. Beyond mixed-initiatively automating certain diagnosis (near Point B), IMPETUS also supports exploration near Point A by enabling pathologists' initial exploration with recommended regions. Image modified based on Blois [33].

evidence (*e.g.*, test results) is gathered until finally a single most probable conclusion can be drawn. According to Blois, AI has been canonically developed to optimize Point B, where a computer program can deterministically confirm whether a patient has a certain disease given all the evidence. As Blois foresaw, a recent development of AI starts to exhibit capabilities towards Point A, *e.g.*, Stanford's CheXpert produces likelihoods of 10+ thoracic diseases based on a chest X-ray image [112]. Similarly, IMPETUS also aims at "reaching Point A" by enabling pathologists' initial exploration with recommended regions.

Overall, IMPETUS provides the first suite of interaction techniques in the medical imaging domain that instantiates mixed-initiative principles [105] for physicians to interact with AI with an adaptive degree of initiatives based on AI's capabilities and limitations. Specifically,

Table 2.1: Spectrum of human and AI initiatives at different AI confidence levels.

AI Confidence	**AI-Initiated Action**	**Physician-Initiated Action**
High	Performing diagnosis automatically in the background; marking WSIs as diagnosed	Doing nothing and accepts AI's results; can re-open a WSI to overwrite AI's result
↑	Pre-filling the diagnosis box without directly labeling the WSI	Performing diagnosis with help from AI predictions; confirming or correcting the pre-filled dialogue
Low	Showing original WSI by default to prompt for manual diagnosis	Performing diagnosis with little input from AI

we focus on the following principles in [105]:

- *Scoping precision of service to match uncertainty.* We first design a rule-based algorithm to identify three levels of uncertainty in AI's performance given a WSI, based on which we then design the corresponding AI-initiated action appropriate for each level of uncertainty (Table 2.1).

- *Providing mechanisms for efficient agent-user collaboration to refine results.* For each AI-initiated action, we also design mechanisms to introduce physician-initiated actions aimed at confirming, refining, or even overriding AI's results (Table 2.1). Further, we extend this principle by leveraging physician-initiated input for 'machine teaching' [184], *i.e.*, an agile labeling technique to dynamically adapt an AI by retraining it with examples of how a physician interpret a patient's histological data.

2.4.2 AI Guiding Pathologists' Attention to Regions of Major Outliers

In our communication with our pathologist collaborator, we learned that one major limitation of pathologists is the ability to efficiently and effectively navigate a large, high-resolution WSI. To address this limitation, we design AI to guide pathologists' attention to regions of major outliers that appear visually different from the rest of the WSI and are more likely to be tumors. Such guidance is manifested in two user interface elements:

(i) **Attention map** visualizes each patch's degree of outlying overlaid on the current WSI (Figure 2.3(a)); *(ii)* **Recommendation boxes** as a more explicit means to draw pathologists' attention to large clusters based on outlier detection results (Figure 2.3(a), yellow box) — these boxes are always visible, whether on the original WSI, on the attention, or on the prediction map (described below).

Implementation When the system is first loaded, a pre-trained InceptionResNetv2 model[1] [111] on PatchCamelyon dataset[2] extracts patch features (patch dimension=$96 \times 96 \times 3$, feature dimension=1536×1) in WSIs (Figure 2.4(a)). Given the imbalance nature of tumor *vs.*normal tissues, in the first iteration, the system performs isolation forest (max_samples=256) [134] outlier detection based on extracted features (Figure 2.4(b)), and the detected outliers are highlighted in the attention map. In the following iterations, the attention map is a combination of outliers (from the initial detection) and high uncertainty patches (from specific models in each iteration)[3]. In order to obtain the recommendation boxes, the system uses a DBSCAN clustering algorithm (min_sample=10, epsilon=3) [70] to cluster WSI patches with attention value (Figure 2.4(c,k)). In order to reduce users' distraction, the recommendation boxes are selected as the two clusters that occupy the largest areas on the WSI in each

[1]We trained this model with image augmentation preprocessing, Adadelta optimizer with initial learning rate 0.1, binary cross entropy loss, 100 iterations with early stopping on validation loss.

[2]`https://patchcamelyon.grand-challenge.org/`

[3]Uncertainty is calculated as Uncertainty $= 1 - |0.5 - \text{Probability}| \times 2$. The attention maps in the following iterations are calculated as the soft-OR of outliers and uncertainty: Attention = Uncertainty $\odot$ Outlier.

iteration.

2.4.3 AI Using Agile Labeling to Train and Adapt Itself On-the-fly

In digital pathology, the main challenge for AI is that, unlike other imaging modalities (*e.g.*, X-ray, CT), histological data (*e.g.*, ovarian carcinoma) tends to have a high variance across slides of different patients (sometimes same patients as well) [122]. Thus a pre-trained model often struggles to generalize to new data. To address this limitation, IMPETUS enables pathologists to use agile labeling to train AI on the fly.

Agile labeling allows a pathologist to directly label on recommendation boxes (Figure 2.1(a)), or to draw a bounding box of tumor-negative patches (Figure 2.1(b)), or a box containing a mix of negative and positive cells, which serve as labels to train an existing model further to incorporate pathologists' domain knowledge. Importantly, such labeling technique is designed to be agilely achievable without incurring significant extra effort that interrupts the main diagnosis workflow.

Implementation Agile labeling does not specifically require users to provide the exact contour of tumor tissues in a WSI, as strongly-supervised learning does. Alternatively, a user can marquee-select a positive box over an area which contains *at least one* tumor patches, or a negative box on *all negative* regions. We implemented a weakly-supervised MIL [230, 17] to learn over such agile labels. To train the model, the system first initializes a *positive_set* and *negative_set*. For each box annotated by a user, IMPETUS first partitions the WSI areas into $96 \times 96 \times 3$ non-overlapping patches, and extracts the feature of each patch by the pre-trained CNN model from Section 2.4.2 ((Figure 2.4(a))). Here, we denote each the extracted feature set as X_i and the box-level annotation from user as Y_i.[4] For a negative box, all the patch features in the box can be included in *negative_set* (Figure 2.4(f)). For positive boxes, the system uses T-SNE [140] to represent the high-dimension

[4]In the MIL setup, each box only has one box-level annotation Y_i.

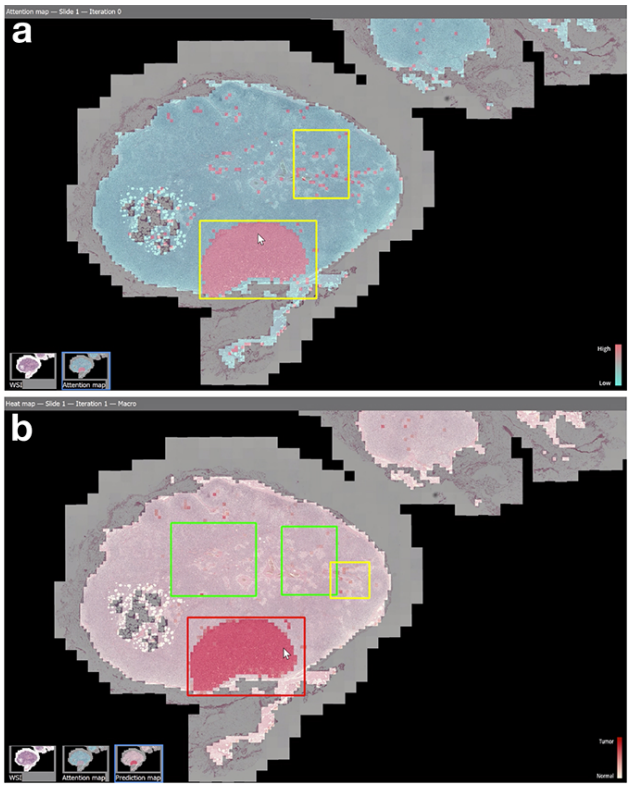


Figure 2.3: The two maps used by IMPETUS to provide guidance and communicate AI results. (a) Attention map, where outlier patches and high uncertainty patches are highlighted in red, while other patches are in blue. The yellow recommendation boxes are generated by clustering attention values. (b) Prediction map, where red shows a high probability of tumor, and white shows a low probability of tumor, as predicted by the AI. The green and red boxes are areas of "normal" and "tumor", as labeled by the pathologist. Recommendation boxes generated by clustering attention values are also visible on this map.

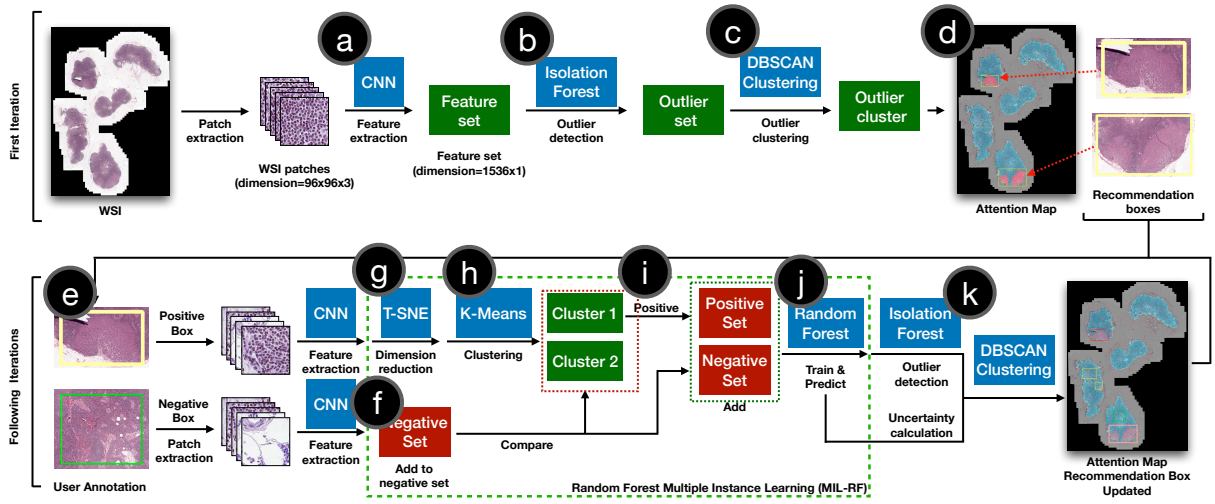


Figure 2.4: Overview of IMPETUS 's AI backend. In the first iteration, IMPETUS first extracts the WSI to non-overlapping patches, followed by (a) feature extraction with a pre-trained CNN (InceptionResNetv2) model; (b) outlier detection by the isolation forest algorithm; (c) outlier clustering with the DBSCAN algorithm. Then, (d) an attention map with outlier clusters and recommendation boxes are generated. In the following iterations, the user first (e) annotates the recommendation boxes with agile labeling. Next, IMPETUS processes negative annotations by (f) adding negative box features to the negative set. For the positive annotation, IMPETUS uses (g) T-SNE to reduce the dimension of positive box features and applies (h) K-Means clustering to split them into two clusters. After that, IMPETUS (i) assigns the two clusters with labels by comparing them to the negative set and only adds features in the positive cluster to the positive set. Last, (j) a random forest classifier learns from the positive and negative set and predicts at a whole-slide level. The attention map and recommendation boxes are generated by (k) clustering from a combination of outliers and uncertain predictions. Procedures between (e - k) are repeated until the doctors are satisfied with the AI performance.

features X_i with two-dimension embedding $\overline{X}_i$[5] (Figure 2.4(g)). Then, K-Means clustering is used to split $\overline{X}_i$ into two clusters: $\overline{X}_i^{(1)}$, $\overline{X}_i^{(2)}$ (Figure 2.4(h)). After clustering, the algorithm compares the two clusters with negative samples from *negative_set* to pick the real positive cluster. After the positive cluster is recognized, all the instances in the positive cluster are included in the *positive_set* (Figure 2.4(i)). Finally, a random forest classifier (MIL-RF, 100 trees, max_depth=100) [37] is trained with the obtained *positive_set* and *negative_set*[6] (Figure 2.4(j)), and the user can continuously provide more annotations until the trained classifier reaches a satisfactory level of performance.

2.4.4 AI Taking Initiatives Appropriately for the Level of Performance Confidence

Even with agile labeling, lightweight on-the-fly learning only has limited improvement compared to training extensively offline. Thus it is crucial to convey the level of AI's performance to the pathologists. In Impetus, AI takes initiatives appropriately for its performance confidence level, as manifested in the following two user interface elements: *(i)* **Prediction map** visualizes current AI's results overlaying the WSI, which serves to inform both the labeling and the usage of the current AI's model (Figure 2.3(b)). *(ii)* **Initiatives based on confidence**—the more uncertain the AI 'feels' about a WSI, the less initiative it takes, as shown in Table 2.1.

Implementation Impetus has a rule-based confidence-level classifier to sort slides into three categories: high-confidence, mid-confidence, and low-confidence. First, predictions of all the patches in the WSI are obtained. A patch has two characteristics: *is_positive* and *is_uncertain*. A patch is positive if the MIL-RF classifier output ¿ 0.5, and is uncertain

[5]The embedding $\overline{X}_i$ is used for clustering for two reasons: *(i)* avoiding K-Means to process high-dimension data, which could prevent clustering performance degradation; *(ii)* better visualizing the high-dimensional embedding space.

[6]The random forest algorithm is used since its "resistance to overfit" [152]. The notion of random forest has been applied to active learning [152], or multiple instance learning algorithms [129].

if MIL-RF classifier output $\in [0.25, 0.75]$. We empirically summarize the confidence-level decision rules[7] as follows:

- If there are more than 200 positive patches AND the number of positive patches is greater than twice the number of uncertain patches, then the slide is predicted as high-confidence;
- Else if there are no outlier clusters, then the slide is predicted as low-confidence;
- Else if the number of uncertain patches is greater than 300, then the slide is predicted as low-confidence;
- Else if the number of positive patches is greater than 200, then the slide is predicted as high-confidence;
- For all other cases, the slide is predicted as mid-confidence.

2.5 Work Sessions with Pathologists

To validate our design of IMPETUS, we observed how pathologists used this tool to perform diagnosis on a clinical dataset [132]. Our goal is to study whether the AI in IMPETUS *(i)* can be compatibly integrated into pathologists' workflow and *(ii)* can provide added values to pathologists' diagnosis process.

2.5.1 Participants

We recruited eight medical professionals from the pathology department in UCLA Health. The participants have experiences ranging from 1 to 43 years, including residents, fellows, and attending pathologists.

[7] ... which can be easily modified as a configuration of our tool.

2.5.2 Test Data and apparatus

We used the Camelyon 17 [132] dataset and selected 16 WSIs[8] that were collected in the same medical center. Participants interacted with IMPETUS on a 15-inch laptop computer using a wired mouse. IMPETUS ran on a Microsoft Windows 10 Operating System using an Nvidia 960M GPU and 16GB RAM.

2.5.3 Design

Our discussion with pathologists collaborators and an initial screening survey indicated that there was not a commonly-used digital pathology tool among the participants. To help pathologists calibrate their experience with IMPETUS, we introduced another tool — ASAP[9], which represents a very basic manual tool for viewing and annotating digital pathology slides. Each pathologist interacted with both IMPETUS and ASAP, which were referred to as System A and System B, respectively, to avoid biasing the pathologists. The order of tools was counterbalanced across the eight pathologists. Twelve of the 16 slides were diagnosed using IMPETUS and the remaining four using ASAP: we chose to keep more slides for IMPETUS as it was the target of our study, whereas ASAP was just to calibrate pathologists' tool experience.

2.5.4 Tasks and Procedure

After briefly introducing the background of computer-assisted diagnosis, we walked each pathologist through a tool and let them practice on a separate toy dataset also gathered from [132]. We then asked questions about how the participant understood different interactive components, whether the tool was easy to learn and use, and whether the tool was

[8]Our pilot studies indicated that 16 is the number of WSIs that would allow us to finish the session in about an hour to most effectively use the pathologists' time.

[9]`https://computationalpathologygroup.github.io/ASAP/`

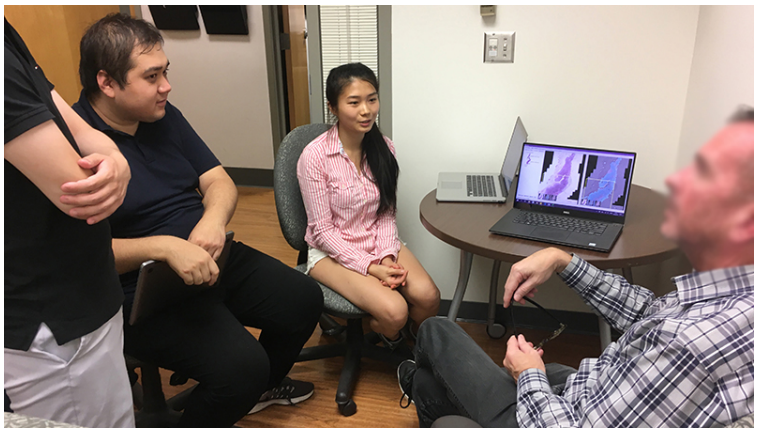


Figure 2.5: We conducted work sessions with eight pathologists from a local medical center to observe how they used Impetus as part of their diagnosis process.

helpful to their diagnosis. Then, the primary task began, which was to diagnose the entire group of WSIs using the provided tool in each condition. A trial started with a participant clicking to open a WSI and finished when they selected a diagnosis and clicked the 'Confirm' button. After each condition, we further conducted a brief semi-structured interview for each participant to summarize their experience, feedback, and suggestions for the tool. Participants took a short break between the two conditions.

2.5.5 Analysis

We employed an iterative open-coding method to analyze the qualitative data collected from the semi-structured interviews with pathologists. Two experimenters coded each participant's data within one day after the study. One experimenter performed the first pass of coding and updated a shared codebook, which was then reviewed by the other experimenter to resolve disagreements. The two experimenters alternated the roles of the first coder and reviewer. After all the participants' data were coded and consolidated, a third experimenter reviewed all the codes and transcripts and resolved disagreements through discussion with the previous two experimenters. Finally, we arrived at six high-level themes, which we

summarize below as lessons learned.

2.6 Findings, Lessons Learned and Design Recommendations

Based on the observations and data from the work sessions with pathologists, we present our findings below, which are summarized into six lessons.

2.6.1 AI Guiding Pathologists' Attention to Regions of Major Outliers

2.6.1.1 Recommendation boxes

(Figure 2.1a) were the most frequently used and discussed features during the study. We observed that in almost all the trials, pathologists started by zooming into the recommendation boxes and tried to provide annotations of the outlined region. Pathologists found it helpful to have such concrete start points in their examination.

> *... [recommendation boxes] narrow down the area of interest ... it helps* (P7)

> *It was less effort because I was focusing only on the attention areas and not focusing on the other areas of the node so it was different from my usual way of looking at a slide.* (P2)

However, pathologists did not always find the recommended regions matched their intuition, and they could not understand why certain regions were recommended.

> *... [the recommendation box] seems a little bit random. It's not necessarily areas that I would [look at] ...* (P5)

> *The things it's focusing on does not correlate with at least what my brain thinks I am looking for.* (P6)

A lack of transparency is not a new problem in recommender system research (*e.g.*, [226]). When introducing IMPETUS, we did explain that recommendation boxes were based on a detection of visual outliers, and all pathologists acknowledged that they understood such a concept. Although such outliers were computed based on histological features (the PatchCamelyon dataset), they did not always agree with what pathologists intuited as 'interesting' regions worth examination. When such a mismatch occurred—*i.e.*, an unexpected case of recommendation, pathologists could no longer reason about the recommendation boxes simply by referring to the abstract concept of 'visual outliers'. At times, pathologists started to develop their own hypothesis of how AI was processing the WSI: "*... it's interesting that it's picking area with fat as area of interest.*" (P2)

Lesson #1 To explain AI's guidance, suggestions and recommendations, the system should go beyond a one-size-fits-all concept and provide instance-specific details that allow a medical user to see evidence that leads to a recommendation.

Recommendation #1: an overview + instance-based explanation of AI's suggestions. Currently, IMPETUS only provides an explanation of the suggested regions at the overview level: a textual description of the outlier detection method as part of the tutorial and visualization (*i.e.*, attention map) that shows the degree of 'outlying' across the WSI. As an addition, we can further incorporate instance-based explanation, *i.e.*, with information specific to a particular patient and a particular region on the patient's slide. The idea is to allow pathologists to question why a specific region is recommended by clicking on the corresponding part of the slide, which prompts the system to show a comparison between the recommended region and a number of samples from non-recommended parts of the slide for the physician to contrast features in these regions extracted by AI. One important consideration is that such an additional explanation should be made available on-demand rather than shown by default, which could defeat the recommendation boxes' purpose to accelerate the pathologists' examination process.

We also found that pathologists wondered what they should do about the area outside of the recommendation boxes:

> *So I just look at the ones in the [recommendation] square?* (P7)

> *Am I supposed to assume the rest of it is normal? I don't have to go searching for the rest of the slides for [tumor]?* (P2)

Pathologists understood the implication *in* the recommendation boxes, *i.e.*, to prioritize certain regions of a WSI and to serve as a 'shortcut' in lieu of scanning the entire WSI. However, it was unclear what was the implication *outside* of the recommendation boxes. This is especially true when pathologists could not find signs of tumor in the recommended regions: the system did not continue to guide them on how to proceed with the rest of the WSI.

Lesson #2 Medical diagnosis is seldom a one-shot task, thus AI's recommendations need to continuously direct a medical user to filter and prioritize a large task space, taking into account new information extracted from a user's up-to-date input.

Recommendation #2: make AI-generated suggestions always available (and constantly evolving) throughout the process of a (manual) examination. For example, in IMPETUS, a straightforward design idea is to show recommendation boxes one after another. We believe this is especially helpful when the pathologist might be drawn to a local, zoomed-in region and neglect looking at the rest of the WSI. The always available recommendation boxes can serve as global anchors that inform pathologists of what might need to be examined elsewhere beyond the current view. This is an example of a multi-shot diagnosis behavior where each shot is an attempt to find tumor cells in a selected region.

2.6.1.2 Attention map

(Figure 2.3a) visualizes outliers detected by the AI — the same information based on which the recommendation boxes were drawn. It was designed to complement recommendation boxes with a backdrop of detailed guidance. We expected pathologists to use the attention map similarly as the recommendation boxes, *i.e.*, to direct their attention to look for more outlying regions for examination. However, pathologists did not find attention map useful:

> *The attention map shows the same thing as the recommended box. The box is enough to direct my attention.* (P2)

> *I don't really see the point of the attention map ... These two maps are redundant.* (P4)

The main difference was that recommendation boxes cost less effort to process, while the attention map needed to be navigated (*i.e.*, panned and zoomed and interpreted (*i.e.*, mentally 'decoding' the color scheme). Further, recommendation boxes provided actionable information (*i.e.*, to look into this box first), while the attention map is action-neutral. Given that pathologists' overall goal is to eliminate the amount of area to study, they tended to prefer less extra effort and information with clearer actionability.

Lesson #3 Medical tasks are often time-critical, thus the benefits of AI's guidance, suggestions and recommendations need to be weighed by the amount of extra efforts incurred and the actionability of the provided information.

Recommendation #3: weigh the amount of extra efforts by co-designing a system with target medical users, as different physicians have different notions of time urgency. Emergency room doctors often deal with urgent cases by making decisions in a matter of seconds, and internists often perform examinations in 15-20 minutes per patient; oncologists or implant specialists might decide on a case via multiple meetings that span

days. There is a sense of timeliness in all these scenarios, but the amount of time that can be budgeted differs from case to case. To address such differences, we further recommend modeling each interactive task in a medical AI system (*i.e.*, how long it might take for the user to perform each task) and providing a mechanism that allows physicians to 'filter out' interactive components that might take too much time (*e.g.*, the attention map in IMPETUS). Importantly, different levels of urgency should be modifiable (perhaps as a one-time setup) by physicians in different specialties.

2.6.2 AI Using Agile Labeling to Train and Adapt Itself On-the-Fly

Prediction map (Figure 2.3b) visualizes current AI's diagnosis of the WSI and was designed to help the pathologists assess the model's performance and decide where they could provide more labels.

However, pathologists used the prediction map differently than we expected. Pathologists would often zoom into recommendation boxes on the WSI, study the region for a few seconds, then switch to the prediction map for a few seconds, and switch back to WSI. They tended to use the prediction map as a tool to help them see if there is something 'interesting' in the current zoomed-in region. Sometimes pathologists used the prediction map for double-checking their developing diagnosis:

> *That was all negative, and I didn't get a strong heatmap signal, so it was confirmatory and somewhat helpful.* (P6)

Interestingly, how pathologists used the prediction map seemed to complement the recommendation boxes: while recommendation boxes told pathologists which region is worth looking at (*i.e.*, might contain tumors), prediction map confirmed pathologists' assumption when they thought a region was of little 'interest' (*i.e.*, no signs of tumor).

Lesson #4 To guide the examination process with prioritization, AI should help a medical user narrow in small regions of a large task space, as well as helping them filter out information within specific regions.

Recommendation #4: use visualization to filter out information, *i.e.*, leverage AI's results to reduce information load for the physicians. An example would be a spotlight effect that darkens parts of a WSI where AI detects little or no tumor cells. Based on our observation that pathologists used AI's results to confirm their examination of the original H&E WSI, such an overt visualization can help them filter out subsets of the WSI patches. Meanwhile, pathologists can also reveal a darkened region if they want to examine further AI's findings (*e.g.*, when they disagree with AI, believing a darkened spot has signs of tumor).

The unexpected usage of the prediction map affected agile labeling, as we discuss below.

Agile labeling (Figure 2.1b) allows a pathologist to label on a recommendation box directly, or to marquee-select a region to coarsely annotate as normal or tumor. In the introduction phase, all pathologists reported having no problem understanding the idea of continuously labeling WSIs to improve the AI:

> *This is actually adding more work for me, but I would be willing to add labels knowing I would be improving the model* (P4)

However, during the tasks, we noticed that almost all the labels were drawn only based on the recommendation boxes. Only one pathologist actively searched for other regions to draw and provide more labels. It seemed that recommendation boxes served as a prompt, and pathologists were unmotivated to label other regions if unprompted.

We believe one fundamental reason is a lack of feedback to inform pathologists how important their labels were to the model retraining. Without such feedback, it might have

been unclear to pathologists whether they needed to provide labels at all, or how much labeling would be enough.

> *Do I need to add labels?* (P6)

> *Should I have provided more labels?* (P5)

We assume that once pathologists see how a prediction map contained inaccurate results, they would be motivated to provide more labels to improve the prediction. However, our observations show that pathologists were more likely to make a diagnosis directly by manual examination, instead of correcting AI's predictions as we expected. Falling back to manual examination seems a more cost-effective alternative to AI automation than than improving the AI iteratively.

Lesson #5 It is possible for medical users to provide labels during their workflow with acceptable extra effort. However, the system should provide explicit feedback on how the model improves as a result, as a way to motivate and guide medical users' future inputs.

Recommendation #5: when adapting the model on-the-fly, show a visualization that indicates the model's performance changes as the physician labels more data. There could be various designs of such information, from showing low-level technical details (*e.g.*, the model's specificity vs. sensitivity), high-level visualization (*e.g.*, charts that plot accuracy over WSIs read) and even actionable items (*e.g.*, 'nudging' the user to label certain classes of data to balance the training set). There are two main factors to consider when evaluating a given design: *(i)* as we observed in our study, whether the design could inform the physician of the model's performance improvement or degradation as they label more data, which can be measured quantitatively as the amount of performance gain divided by the amount of labeling work done; *(ii)* as we noted in Lesson #2, whether consuming the extra information incurs too much effort and slows down the agile labeling process, and whether there is actionability given the extra information about model performance changes.

2.6.3 AI Taking Initiatives Appropriately for the Level of Performance Confidence

As shown in Table 2.1, AI's level of initiative is mediated based on its level of confidence about the model's performance. For low-confidence cases, AI took no initiative, and all pathologists were mostly unaware of AI's presence, while they simply focused on performing the usual manual diagnosis. For high-confidence cases, as expected, pathologists quickly confirmed AI's proactive diagnosis of macro — the easiest type of tumor to detect by both pathologists and AI. However, when it comes to cases diagnosed as negative by the AI, pathologists tended to perform a manual diagnosis anyway:

> *On the ones that it said it's confident but didn't really tell you it's negative, I still felt like I had to look at those to confirm. I wasn't going to trust the system [to confirm] that it's negative* (P2)

In pathology, in order to rule out tumors, pathologists have to thoroughly examine the entire WSI, whereas it only takes one positive case to diagnose the lymph node as positive. Thus there was a discrepancy of trust between macro vs. negative, despite that AI treats both equally as different labels of a slide image and categorizes both as high confidence.

Lesson #6 Tasks treated equally by an AI might carry different weights to a medical user. Thus for medically high-staked tasks, AI should provide information to validate its confidence level.

Recommendation #6: provide additional justification for a negative diagnosis of a high-staked disease. For example, when IMPETUS concludes a case as negative, the system can still display the top five regions wherein AI finds the most likely signs of tumor (albeit below a threshold of positivity). In this way, even if the result turned out to be a false negative, the physicians would be guided to examine regions where the actual tumor cells are

likely to appear. Beyond such intrinsic details, it is also possible to retrieve extrinsic information, *e.g.*, prevalence of the disease given the patient's population, or similar histological images for comparison. As suggested in [217], such extrinsic justification can complement the explanation of a model's intrinsic process, thus allowing physicians to understand AI's decision more comprehensively.

For the mid-confidence case, AI was designed to pre-fill the diagnosis dialog (but without any confirmative action) as a way to hint its prediction without signaling any conclusive decision. This design did not seem to have noticeable effects on the pathologists, which echos Lesson #3 that information needs to present actionability in order to affect a medical user's workflow.

2.7 Chapter Summary

In this chapter, we first propose a series of collaborative techniques to engage human pathologists with AI given AI's capabilities and limitations, based on which we prototype Impetus — a tool where an AI takes various degrees of initiatives to provide various forms of assistance to a pathologist in detecting tumors from histological slides. We summarize observations and lessons learned from a study with eight pathologists and discuss recommendations for future work on human-centered medical AI systems.

CHAPTER 3

Navigating Challenging Pathology Examinations with Human-AI Collaboration

This chapter is based in part on the following publication:

Hongyan Gu, Chunxu Yang, Mohammad Haeri, Jing Wang, Shirley Tang, Wenzhong Yan, Shujin He, Christopher Kazu Williams, Shino Magaki, and Xiang 'Anthony' Chen. "Augmenting pathologists with NaviPath: design and evaluation of a human-AI collaborative navigation system." In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pp. 1-19. 2023.

3.1 Introduction

One crucial step of cancer diagnoses is the pathologists' examinations of tumors through an optical microscope. With the recent development of digital pathology [73, 157], tumor specimens can be scanned into high-resolution digital scans, allowing medical professionals to access, analyze, and share these scans with digital interfaces [143, 95, 175]. However, literature has suggested that it might take longer for pathologists to examine digital scans compared to when using microscopes [198, 99]. The main culprit is the difficulty in navigation — pathology scans usually have extremely high resolutions ($(\sim 10^6)^2$ pixels) compared to commercial off-the-shelf computer displays ($\sim 8.3 \times 10^6$ pixels for 4K UHD resolution). Therefore, pathologists are required to frequently manipulate (*i.e.*, zooming, panning) the viewport to gather necessary information for diagnoses [174].

Research has long realized the difficulty in navigating high-resolution images and proposed various interface designs to assist users with general navigation tasks (*e.g.*, map exploration) [25, 224, 94, 177, 57]. However, we believe necessary adaptations should be considered to enable seamless integration into pathologists' workflows, because of three problems in human navigation of pathology scans: *(i)* pathologists' navigation is usually substantially complicated because some pathology patterns (*e.g.*, mitosis in low-grade meningiomas [137]) have a low prevalence rate (<100/scan) and have extremely small dimensions compared to pathology scans (ratio up to 1:2000) [12]; *(ii)* pathologists require specific domain knowledge and navigation strategies [174, 149] to facilitate their examinations, which current navigation systems for general use rarely consider; *(iii)* although AI can be used to accelerate navigation, the lack of consideration towards integrating AI into pathologists' workflows might discourage them from using human-AI systems in practice, as suggested in previous studies [222]. Fortunately, recent HCI-AI-Health works have demonstrated prototypes and designs to close the gap between medical professionals and AI, which has facilitated human-AI communication and was viable to improve doctors' works in various medical application domains, such as general medicine [221, 127, 178], radiology [48, 47] and pathology [43, 113]. Motivated by the success of these advancements, this chapter continues to build integrable systems by taking doctors' domain knowledge into account, with a focus on supporting the navigation process in pathology.

To this end, we conducted a formative study with six medical professionals in pathology from two medical centers to enrich our understanding of their navigation processes. Specifically, we observed how they navigated pathology scans to search for mitoses[1], a critical pathology pattern that relates to cancer malignancy and patient prognosis [61]. We summarized three observations that cross-validate the findings in previous research [183, 84, 174, 149]:

[1]The mitosis is selected because *(i)* the size of mitoses is small ($\sim 10\mu m$) compared to the size of pathology scans; *(ii)* the prevalence of mitoses is low ($< 0.2/(1,600)^2$ pixels in specific carcinomas) [12].

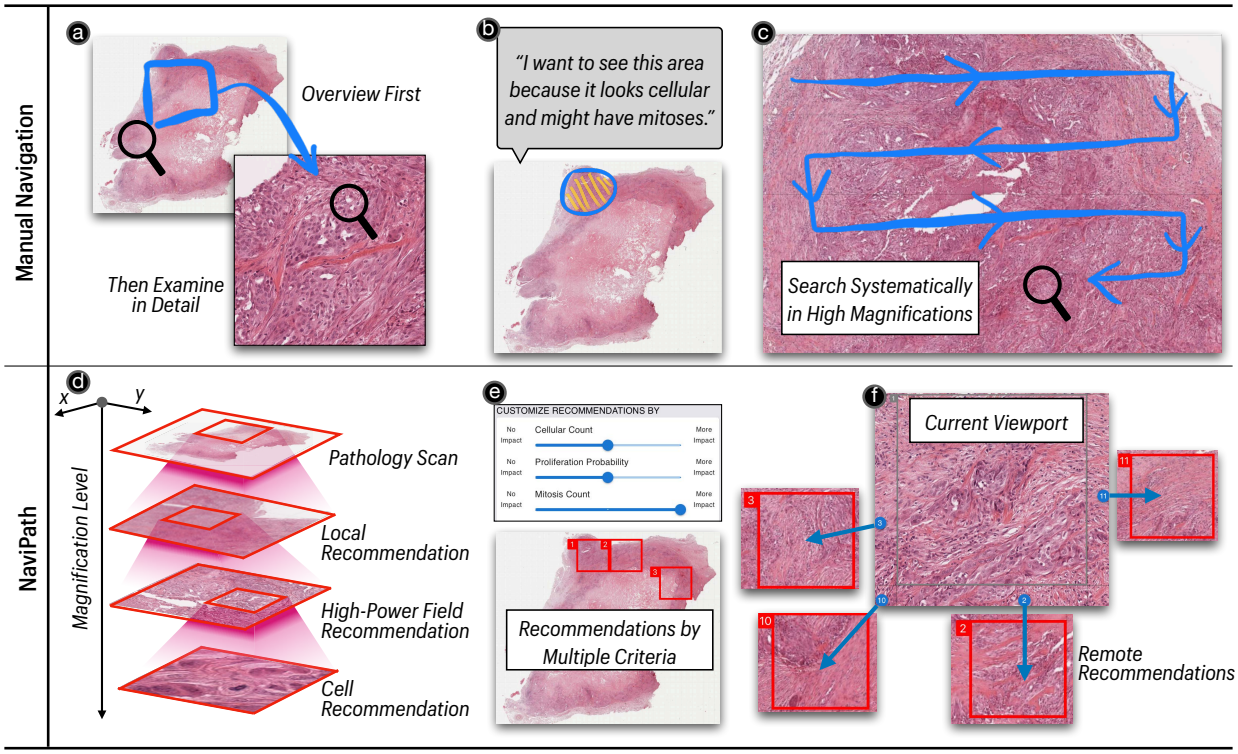


Figure 3.1: Comparison between pathologists' manual navigation in practice *vs.*NaviPath's designs. Observations on pathologists' manual navigation: (a) Pathologists usually overview a pathology scan with low magnifications, followed by switching to higher magnifications to examine regions of interest in detail; (b) Pathologists might refer to macroscopic patterns to locate ROIs in the low magnification; (c) Pathologists employ a systematical searching strategy in high magnifications. NaviPath's designs: (d) NaviPath harnesses AI to generate hierarchical "Local", "High-Power Field", and "Cell" recommendations, covering multiple magnification levels; (e) NaviPath utilizes AI to calculate three criteria that pathologists usually consider to generate recommendations; (f) Once in high magnifications, NaviPath places navigation cues on the edge of the interface, enabling pathologists to jump to remote AI recommendations without manual panning.

1. **Overview first, then detail**: Pathologists followed this pattern of interacting with visual data as found in earlier works [183, 84]: they started with an overview of the scan using low magnification, then selected a few **regions of interest (ROIs)** and studied each ROI in detail using higher magnifications (see Figure 3.1(a));

2. **Using macroscopic patterns to locate ROIs in the low magnifications**: Pathologists referred to macroscopic patterns visible in low magnifications that were associated with occurrences of mitoses (see Figure 3.1(b)) to locate ROIs in low magnifications;

3. **Low throughput in high magnifications**: Pathologists adopted a cautious and comprehensive navigation strategy (see Figure 3.1(c)) [149] to avoid missing crucial pathology patterns, causing low throughput under high magnifications.

After accumulating the empirical evidence to verify existing knowledge in pathologists' navigation, we designed NaviPath — a human-AI collaborative navigation system that bridges the gap between AI and pathologists by integrating doctors' domain knowledge. Currently, we focus on pathologists' practices of examining mitosis as a showcase for NaviPath. Mirroring the three observations mentioned above, we propose three design components of NaviPath:

1. **Hierarchical AI Recommendations**: As shown in Figure 3.1(d), NaviPath employs AI to generate hierarchical recommendations across multiple magnification levels to support pathologists' "overview first, then detail" workflows. Specifically, the "Local" recommendation helps pathologists to quickly focus on a rough interest area in low magnification; the "High-Power Field" recommendation allows pathologists to narrow down and examine in detail using a median magnification level; and the "Cell" recommendation assists pathologists in adjudicating whether a suspected cell is mitotic in the highest magnification.

2. **Customizable Recommendations by Multiple Criteria**: NaviPath generates hierarchical AI recommendations with three criteria that pathologists usually consider to localize ROIs in practice (*i.e.*, cellular count, proliferation probability, and mitosis count). Furthermore, NaviPath permits pathologists to customize AI recommendations according to their examination preferences by a group of slide-bars (Figure 3.1(e), top figure).

3. **Cue-Based Navigation for High Magnifications**: To cope with pathologists' low throughput under high magnifications, NaviPath adapts the notion of existing cue-based navigation designs [224] and places short-cut navigation cues on the edge of the viewport (Figure 3.1(f)). This design enables users to jump to remote AI recommendations without manual panning, which can improve pathologists' navigation efficiency.

We recruited 15 medical professionals in pathology from five medical centers across two countries to validate NaviPath. We discovered that, compared to traditional manual navigation:

1. Participants' navigation efficiencies were significantly improved (p=0.002, r=0.579, from Wilcoxon rank-sum test) with NaviPath: they saw more than twice the number of the target pathology pattern (*i.e.*, mitosis) in unit time on average;

2. Both participants' precision and recall on identifying the target pathology pattern were significantly improved (precision: p <0.001, recall: p <0.001, from post-hoc Dunn's test) with NaviPath. Meanwhile, compared to the AI, participants' average recall and precision were improved by 20.21% and 21.51% by NaviPath, respectively;

3. Participants reported significantly less mental effort (p <0.001, r=0.658, from Wilcoxon rank-sum test, same following), had higher confidence (p=0.004, r=0.530), and were

more likely to use NaviPath in the future (p=0.001, r=0.594), based on a post-study questionnaire.

3.1.1 Contributions

We propose and validate the implementation of an AI-assisted tool in pathology, NaviPath, to enhance the navigation for pathologists by incorporating domain knowledge and considering workflow integration in practice. NaviPath could reduce pathologists' burdens by automating navigation with an AI-assisted algorithm while its collaborative workflow augments pathologists' work. Throughout a user evaluation study with medical professionals, we demonstrated that our human + AI system could improve doctors' navigation efficiencies and lead to a higher examination quality. Instead of imposing an end-to-end, black-box AI into their workflows, this chapter closes the gap between medical professionals and AI by embedding doctors' domain knowledge and enabling them to delegate tasks to AI according to their preferences. Although majorly focused on mitosis in pathology, we further provide design insights for HCI researchers on how AI and medical professionals can work collaboratively to support medical decision-making in light of our observations in the evaluation study.

3.2 Task Design and Medical Background

3.2.1 Task Selection and Generalizability of the Task

This chapter selects the task of mitosis (a type of histology pattern) detection in brain tumors of meningiomas (Figure 3.2(a)). The significance of mitosis stems from its critical role in tumor assessment and patient management for meningiomas [61, 137, 85]. Despite their importance, pathologists' evaluation of mitoses often faces substantial difficulties. The intricacies lie in mitotic figures' small size, low prevalence, and heterogeneous distribution

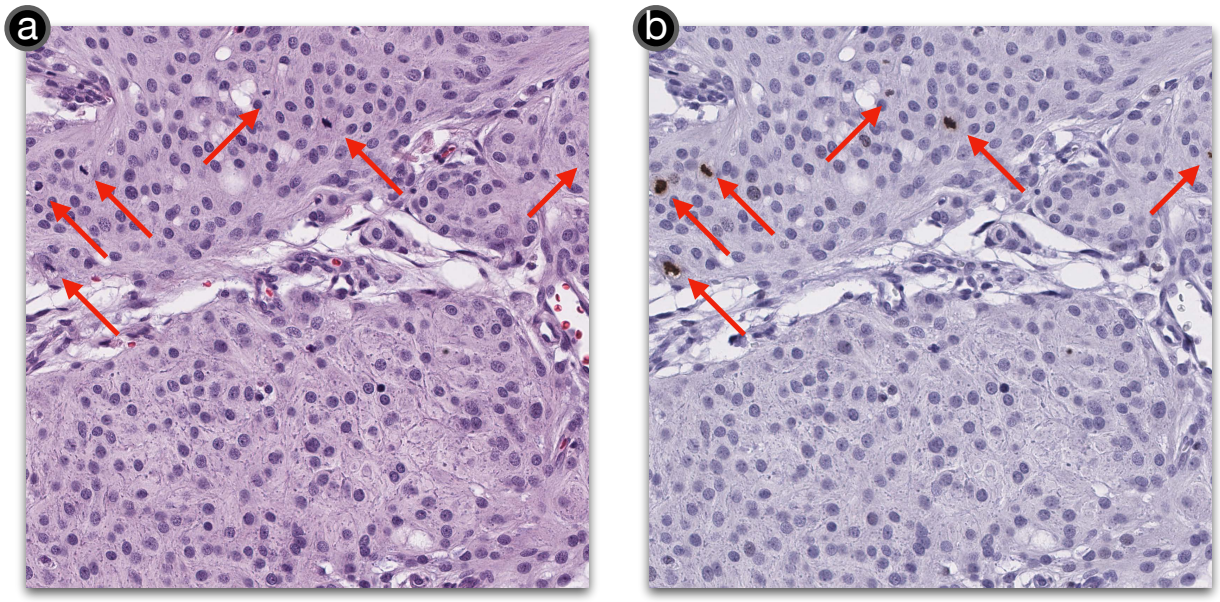


Figure 3.2: (a) An example region-of-interest image used in the user study, with arrows pointing at the ground truth mitoses; (b) The anti-body test used by the three doctors to annotate the ground truth mitoses. Mitoses were shown in brown (as pointed by the arrows) in the anti-body test.

[12, 29]. These complexities contribute to low reported sensitivities, consistencies among pathologists, and examination efficiencies for mitosis evaluation [58, 146, 205, 92], which could negatively impact medical outcomes.

According to the 2021 World Health Organization central nervous system tumor classification guidelines, mitosis serves as a critical diagnostic criterion for grading numerous brain tumors, such as IDH-mutant astrocytoma, oligodendroglioma, and ependymoma [137]. Going beyond mitoses, pathologists may also be required to detect small-scale, sparsely distributed patterns in large scans, such as finding small tumor deposits within lymph nodes in breast cancer or malignant melanoma [168]. In a more general context, similar visual search tasks also exist in high-stakes domains where AI assistance could be valuable. For instance, security personnel must swiftly identify potential threats like explosives in X-ray

Table 3.1: Demographic information of the participants in the formative study.

ID	Occupation	Years of Experience	Medical Center
FP1	Resident	4	MC3
FP2	Resident	4	MC3
FP3	Attending	10	MC1
FP4	Resident	4	MC1
FP5	Resident	4	MC3
FP6	Resident	3	MC3

scans [216], and emergency responders rely on timely assessments of disaster impacts from satellite imagery [151].

3.3 Formative Study and System Requirements

We conducted a formative study with six medical professionals in pathology (referred to as FP1 – FP6) from two medical centers to study how pathologists examine digital scans for mitosis evaluation (see Table 3.1 demographic information of participants). The participants were recruited using flyers sent in mailing lists and word-of-mouth. For each participant, we first introduced the mission of the project. Then, we presented a pathology scan selected from [12], and asked participants to assess the activity of mitosis (a pathological pattern). We followed up with a semi-structured interview and inquired how they navigated the scan to find mitoses. Finally, we presented a series of candidate mock-ups of NaviPath and collected participant feedback. The length of the semi-structured interview was about 30 minutes, and the average duration of each study was about 60 minutes.

3.3.1 Observations

We analyzed the transcribed interview recording using the following approach: first, two researchers summarized the observations individually; then, a third researcher reviewed the observations and addressed the disagreements. We concluded three observations of how pathologists navigate pathology scans (without AI) in their practice, which cross-validated findings from previous work on humans' navigation patterns in high-dimensional visual data.

- **O1: Overview first, then detail.** To search for mitoses, pathologists would first stay in low magnifications to get an overview of the scan, then select a few ROIs and study each ROI in greater detail using higher magnifications. Such a routine was also described in previous works in the general domain of information searching [183, 84] and pathology [174]. Pathologists adapted the searching strategy because of the size difference between mitoses and pathology scans — mitosis is a small-sized pathology feature and can hardly be observed without high magnifications (*i.e.*, $\sim$ $\times$400 magnification). However, scanning the entire slide systematically in $\times$400 [149] can be substantially time-consuming because the field of view under $\times$400 is small compared to the pathology scan: a field of view under $\times$400 has a size of $0.16mm^2$, while a typical $\times$400 pathology scan usually has a size of $\sim 100mm^2$. In our study, all six participants searched for mitoses more efficiently: first, they rapidly covered the scan in low magnifications ($< \times$50) as an overview. After that, they selected a few ROIs to proceed: for each ROI, they switched to medium-magnification ($\sim \times$200) to maximize their fields-of-view while preserving cellular details. If a suspected cell was found, they would dive into high-magnification ($\times$400) and make an adjudication.

- **O2: Using macroscopic patterns to locate ROIs in the low magnification.** To locate the mitosis, pathologists used not only the microscopic features (only visible in $\times$400) but also referred to macroscopic patterns (visible even in $< \times$50) that were associated with the occurrences of mitoses. Specifically, pathologists located ROIs in

low-magnification by evaluating the cell density — *"if it (an ROI) is more cellular, it is more likely to have mitoses"*(FP3).

- **O3: Low throughput in higher magnifications.** While pathologists relied on the cell density to select ROIs from low magnifications, they were likely to 'get lost' once they had switched to higher magnifications. This is because there was a lack of visual landmarks under high magnifications in tumor scans (*i.e.*, the 'desert fog' problem [118]). From the study, we observed that some participants preferred to use a cautious and comprehensive navigation strategy [149] (see Figure 3.1(c)) to avoid missing critical findings that might overturn the diagnosis. However, because not all areas under the high magnifications include mitoses, the navigation strategy might be less efficient and more prone to causing fatigue.

3.3.2 System Requirements

Based on our observations, we propose the following three system requirements for human-AI navigation systems for pathologists:

- **R1: Covering multiple magnification levels**. In accordance with pathologists' "overview first, then detail" navigation processes, the system should provide AI support across multiple magnification levels. For example, recommendations in low magnifications can draw pathologists' attention by pointing out rough areas of interest, while those in higher magnifications should offer more precise guidance in locating ROIs.
- **R2: Incorporating pathologists' domain knowledge**. To bridge the gap between pathologists and AI, instead of employing end-to-end, black-box AI, the system should adapt AI closely to pathologists' domain knowledge and involve criteria that pathologists use in practice to generate AI recommendations. Moreover, because pathologists might have diverse preferences and AI can be imperfect [187, 11], the system should

allow users to customize AI recommendations by emphasizing or ruling-out specific criteria.

- **R3: Accelerating navigation in high magnifications**. To address the low-throughput issue, the system should offer interface designs that enable users to navigate efficiently among the AI recommendations in high magnifications, without getting lost.

3.4 Design of NaviPath

In this section, we first introduce four design components used in NaviPath. We then describe how NaviPath augments pathologists' navigation by describing an example workflow.

3.4.1 Design Components

Corresponding to the three system requirements, we propose three key designs in NaviPath: **Hierarchical AI Recommendations**, **Customizable Recommendations by Multiple Criteria**, and **Cue-Based Navigation for High Magnifications**. Furthermore, we employ the design of **Explaining Each Recommendation** to help pathologists comprehend AI findings.

3.4.1.1 Hierarchical AI Recommendations

Following pathologists' navigation processes for mitosis searching, NaviPath offers AI recommendations of three sizes[2] to provide assistance across multiple magnification levels (system requirement **R1**):

1. The "Local" recommendation (size=10,080×10,080 pixels[3]) simulates pathologists'

[2]Specific sizes were justified by consulting with a board-certified pathologist (experience = 10 years)

[3]The size of one pixel is 0.25μm.

overviewing processes in low magnification. As shown in Figure 3.3(a), the recommendations are red boxes visible in the pathology scan without zooming. Local recommendations can provide rough directional guidance for pathologists; users can prioritize their examination on AI-selected regions without evaluating the scan manually.

2. There are multiple **"High-Power Field" (HPF)** recommendations (size=1,680×1,680 pixels) within a Local recommendation (Figure 3.3(b), red boxes). The HPF recommendation gives more precise ROIs at a higher magnification level, allowing users to examine them in detail. It has the same field of view as ×400 in optical microscopes that pathologists use in practice, freeing them from spending extra effort on adapting to the digital interface.

3. The "Cell" recommendation (size=240×240, Figure 3.3(d)) points out the most precise location of each suspected mitosis reported by AI. It augments pathologists' mitosis evaluations by transforming a visual search task (*i.e.*, finding where mitoses are) into the adjudication (*i.e.*, whether a Cell recommendation includes mitosis).

For all three levels, users can select a recommendation by double-clicking on it, and NaviPath will automatically zoom and center the viewport to the selected recommendation. Therefore, with hierarchical AI recommendations, users can proceed through magnification levels by selecting recommendations on the next level (*e.g.*, Figure 3.3(a)→(b), (b)→(c), (c)→(d)). Users may ignore the recommendation if an undesired one appears.

3.4.1.2 Customizable Recommendations by Multiple Criteria

NaviPath embeds pathologists' domain knowledge and employs three deep learning models (Figure 3.4(c)) to calculate three criteria for obtaining Local and HPF recommendations (system requirement **R2**):

1. **Cellular Count**: Similar to how pathologists leverage the cell density to locate ROIs

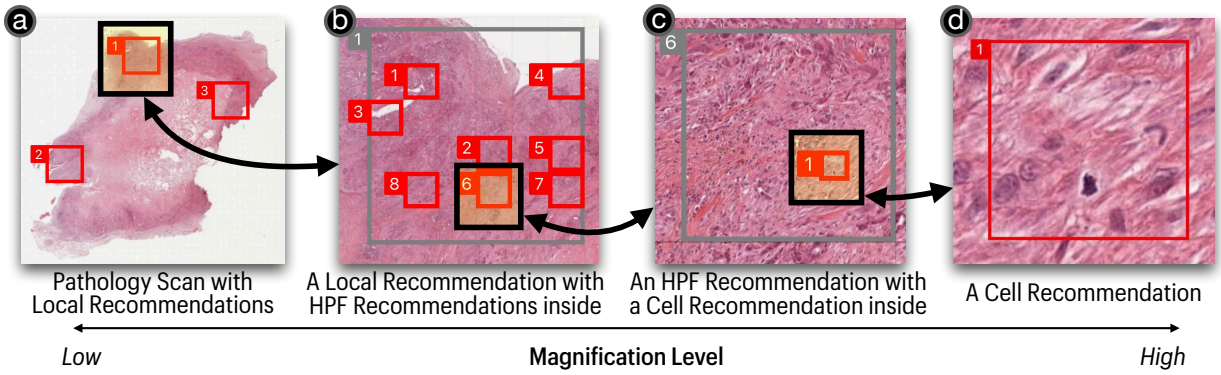


Figure 3.3: NaviPath generates hierarchical AI recommendations across multiple magnification levels: (a) Local recommendations (red boxes) lie in the lowest magnification, and can be seen directly on the pathology scan without zooming; (b) there are multiple High-Power Field (HPF) recommendations (red boxes) inside one Local recommendation (gray box); (c) once in an HPF recommendation (the gray box), users can select and see (d) a Cell recommendation with the highest magnification.

in the low magnification, NaviPath employs a state-of-the-art nuclei segmentation model (*i.e.*, HoVer-Net) to count cell numbers and capture cellular areas from the pathology scan.

2. **Proliferation Probability**: Mimicking pathologists' judgements of whether an area needs further attention in $\times$400 from $\times$200 views, NaviPath uses an EfficientNet-b3 model [193] to predict the proliferation probability — a criterion that relates to whether an ROI is likely to include mitosis, based on AI's impressions from $\times$200 magnification.

3. **Mitosis Count**: Corresponding to pathologists' mitoses searching in $\times$400, NaviPath utilizes a classification model (*i.e.*, EfficientNet-b3) to detect mitotic figures from the highest magnification.

As for Cell recommendations, NaviPath directly pulls the positive results from the mitosis AI and visualizes them on the interface.

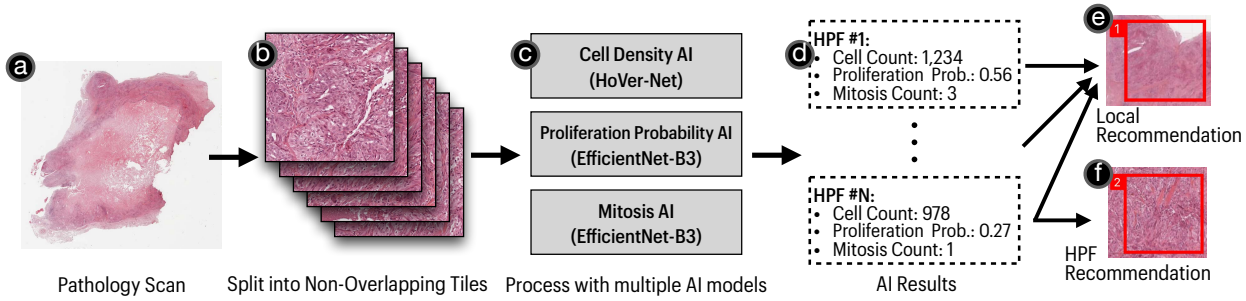


Figure 3.4: Generating Local and HPF recommendations with multiple criteria: (a) a pathology scan is first (b) split into non-overlapping tiles. Then, NaviPath uses (c) three AI models to analyze each tile to obtain (d) scores of cellular count, proliferation probability, and mitosis count. NaviPath will (e) aggregate scores from multiple tiles to generate Local recommendations, or (f) directly use these scores for HPF recommendations.

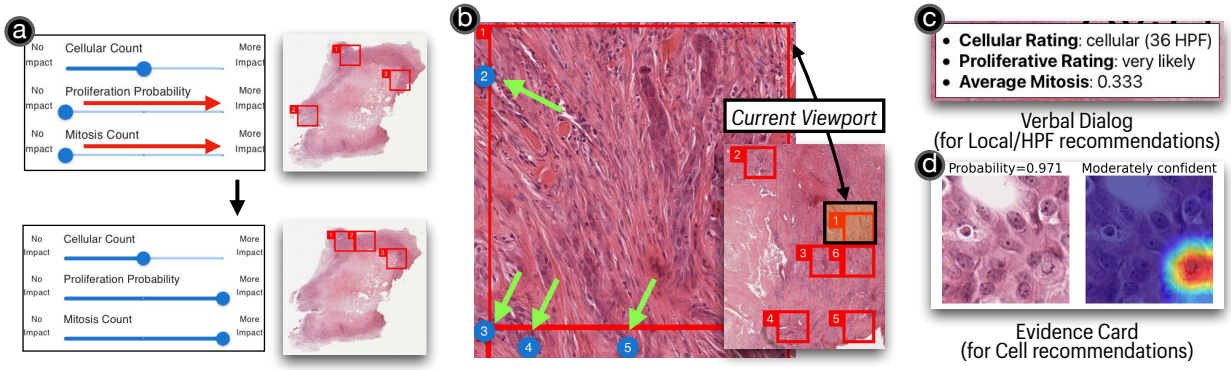


Figure 3.5: (a) NaviPath supports users to customize AI recommendations with a group of slide-bars: users can emphasize or rule out each of the three criteria (*i.e.*, cellular count, proliferation probability, mitosis count) for NaviPath's recommendations; (b) NaviPath places navigation cues (pointed by arrows) that enable users to hop to remote recommendations. The figure on the right provides an overview of off-screen recommendations; (c) An example of NaviPath's verbal dialog explanation for Local/HPF recommendations; (d) An example of the explanation card for NaviPath's Cell recommendations.

Since pathologists might use the three criteria differently in practice, NaviPath supports users to customize AI recommendations by emphasizing or ruling out specific criteria with a group of slide-bars, as shown in Figure 3.5(a). For example, giving the "Proliferation Probability" and "Mitosis Count" higher weight by moving the slide-bar to the right will force NaviPath's recommendations to lean on these criteria. NaviPath will then re-calculate and update recommendations based on the user's input. What's more, users can also adjust the sensitivity of recommendations. For example, if users wish to see more recommendations, they could tune up the "Mitosis Sensitivity" slide-bar (see Figure 3.6(f), the fourth slide-bar).

NaviPath ranks all recommendations according to the current customization setting. Based on the ranking result, it assigns each AI recommendation an index (*e.g.*, Figure 3.6(a), the number on the top-left corner of the recommendation). The smaller the index, the greater the importance and need to be examined with high priority. The index number gives users "actionable" advice [89] and can help them focus on critical areas in limited time. Please refer to the Section 3.5.1 for the implementation details of AI models and the recommendation ranking algorithm.

3.4.1.3 Improving Navigation in High Magnifications

Following system requirement **R3**, NaviPath uses two designs to optimize pathologists' navigation in high magnifications:

First, NaviPath enables pathologists to pan discretely in high magnifications. Specifically, after examining each HPF recommendation, users can double-click on the screen's edge to pan discretely to an adjacent one. Compared to the conventional manual panning with mouse-dragging, this design can accelerate users' interaction speeds: according to Fitt's Law [75], screen edges have infinite width, so it follows that.

Moreover, to increase pathologists' efficiency in seeing remote recommendations, our proposed system adapts the notion of citylight [224] by placing navigation cues on the edge of

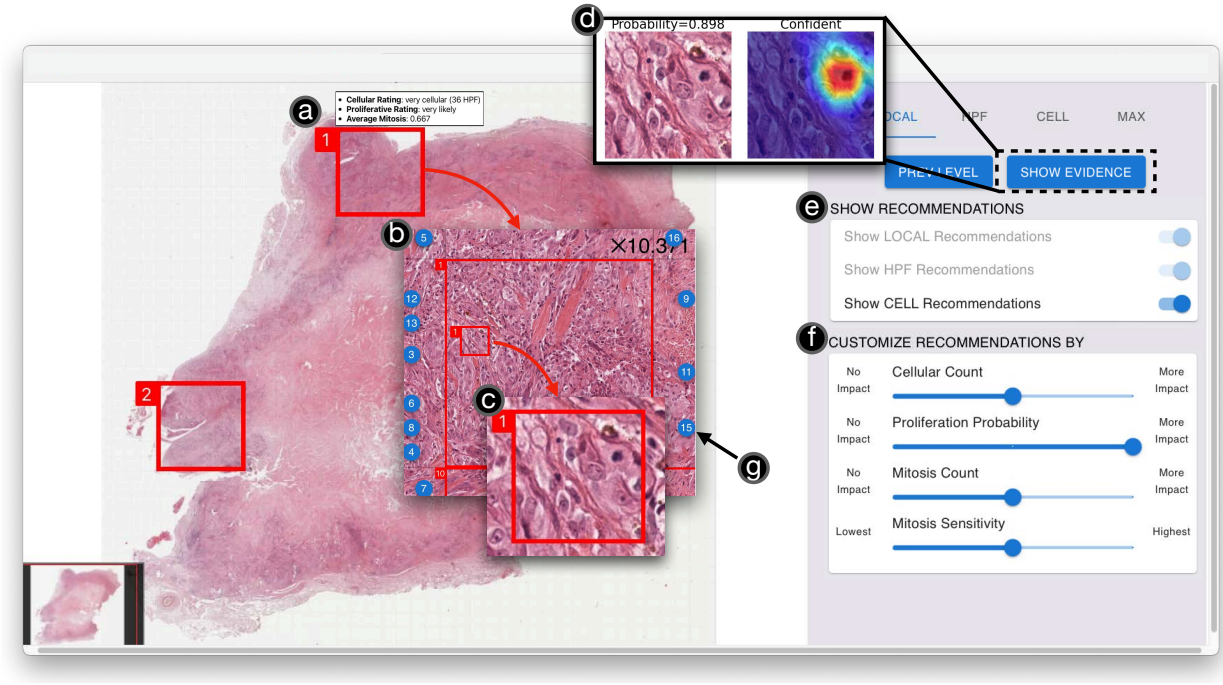


Figure 3.6: Overview of NaviPath's interface. (a) A Local recommendation (red box) with an explanation dialog. The number on the top-left corner represents the index of the recommendation (same for HPF and Cell recommendations); (b) An example of an HPF recommendation; (c) An example of a Cell recommendation; (d) An explanation card for a Cell recommendation, including the AI probability, confidence level, and a saliency map; (e) Users can switch on and see each level of recommendations on-demand; (f) Users can customize the recommendations with a group of slide-bars; (g) A navigation cue that allows users to jump to a remote recommendation. The number indicates the index of the remote recommendation.

the interface (Figure 3.5(b), pointed by arrows). The location of the navigation cue indicates the relative direction between the remote HPF recommendation and the current viewport, while the number represents the ranked index of each recommendation. With navigation cues, users can become aware of the spatial distribution and importance of off-screen targets. They can also click on navigation cues to hop to remote HPF recommendations without manual panning.

3.4.1.4 Explaining Each Recommendation

Since one criticism of deep learning models in pathology is that there is a lack of interpretability [189], **explainable AI (XAI)** techniques have been utilized to make AI "transparent, understandable and reliable" to pathologist users [162]. In NaviPath, we followed the suggestions from [89] and attached an explanation for each AI recommendation. Specifically, for Local and HPF recommendations, NaviPath presents users with a verbal dialog, which includes qualitative descriptions of AI results on the cellular count, proliferation probability, and mitosis count (Figure 3.5(c)). The dialog helps users decide whether they should select and study recommended areas. Moreover, NaviPath explains each Cell recommendation with an explanation card (Figure 3.5(d)). The explanation card demonstrates the classification probability, the confidence level, and a saliency map for a positive mitosis classification result, which provides information from AI's perspective to assist pathologists' mitosis adjudications. Detailed procedures of explanation generation are described in the Section 3.5.1.

3.4.2 Navigating with NaviPath

A typical page of NaviPath is shown in Figure 3.6. A user's workflow in NaviPath starts by switching on (Figure 3.6e) and seeing Local recommendations (Figure 3.6a). The number on the top-left corner of each recommendation box is the ranking index, and users may view recommendations by ascending index order. In each Local recommendation, users

can continue to drill down and see HPF recommendations (Figure 3.6b). In each HPF recommendation, users can continue to see Cell recommendations (Figure 3.6c) that show the precise locations of detected mitoses. For each Cell recommendation, users can view an explanation card on-demand (Figure 3.6d). After examining each HPF recommendation, users may click on the numbered navigation cue (Figure 3.6g) to hop to a remote HPF recommendation. During users' examination, they may customize the recommendations by interacting with a group of slide-bars (Figure 3.6f). Users' workflow ends when they are confident of signing out the case.

3.5 Implementation of NaviPath's Data Processing Pipeline

3.5.1 Use Multiple AI Models to Calculate Multiple Criteria

We first split each scan into non-overlapping tiles. Each tile has a size of one High-Power Field (400×, size=1680×1680 pixel[4]). For each tile, we first applied a Hover-Net model [86] to obtain the cellular count. Second, to estimate the proliferation probability, we trained an EfficientNet-b3 classification model [193] with a dataset with 100× snapshots of HPFs (*i.e.*, $1680 \times 1680 \rightarrow 420 \times 420$ pixel) as X (input), and corresponding binary labels (*i.e.*, 0: does not have mitosis in HPF, 1: has mitosis in HPF) as y (output). Third, we trained another binary classifier to detect mitosis under 400× with an EfficientNet-b3 model similar to the procedure as described in [88], where its size of input was 240×240. We then applied the mitosis model with a sliding window technique (step size=60 pixel), and further used the non-max suppression as post-processing method to eliminate overlapping boxes. After the scores of each tile were calculated, NaviPath can directly use them for HPF recommendations. As for Local recommendations, NaviPath averages the HPF scores within the Local area for recommendations.

[4]The dimension of one pixel is $0.2500\mu m$ unless specified.

3.5.2 Generate Explanations for Each Recommendation

For the verbal dialog for Local and HPF recommendations, we formulated a rule to generate the descriptions of three criteria reported by AI. From the formative study, we learned that users might encounter difficulties in interpreting the values of cellular count and proliferation probability. Therefore, we converted these two scores into five-scale descriptors according to their value percentiles (*i.e.*, *"{very/·/moderately/slightly/not} likely"*). We also included the mitosis count in the verbal dialog to inform users of the mitotic activity in recommended areas. As for the explanation card for each mitosis detection, NaviPath selected and extracted the last feature map (25^{th} layer) of the mitosis model and generates the class activation map (saliency map) with the GradCAM++ [52]. Furthermore, NaviPath calculated the uncertainty score with Equation 3.1, where the Bayesian uncertainty is the standard deviation of 50 Dropout-enabled predictions with the recommended cell image as input [79], and the noise uncertainty stands for the standard deviation of model predictions on 50 noise-augmented cell images as input [15]. Similar to the verbal dialog, the uncertainty scores were also converted to five-scale descriptors according to their value percentiles.

$$\text{Uncertainty} = \sqrt{\text{Bayesian Uncertainty}^2 + \text{Noise Uncertainty}^2} \tag{3.1}$$

3.5.3 Generate and Rank Recommendations

The ranking for a Local recommendation was obtained by Equation 3.2. Specifically, for the r^{th}, s^{th} HPF that is inside the local recommendation, $C_{r,s}$ stands for the AI-reported cellular count, $P^{\text{prof.}}_{r,s}$ is the proliferation probability, and $\sum \mathbb{1}[P^{\text{mitosis}}_{r,s} \geq 0.66]$ indicates the number of mitosis detected with the highest sensitivity. Note that the cellular count scores and the mitosis count were normalized to $[0, 1]$ by dividing with the max scores in the subset (*i.e.*, 5009, 4). Subsequently, users can control the three weights (*i.e.*, $W_{\text{cellular\_count}}, W_{\text{prof.}}, W_{\text{mitosis}}$) with the first three slide-bars in NaviPath. In each Local recommendation, NaviPath ranks the

scores of all candidate HPFs ($S_{i,j}^{\text{HPF}}$) within the Local Recommendation territory by Equation 3.3, where $\theta_{\text{Sensitivity}}$ can be adjusted by the user input. In each HPF recommendation, NaviPath ranks Cell recommendation scores according to Equation 3.4, where U_{mitosis} is the uncertainty score of each detection (calculated by Equation 3.1). Specific weights, thresholds and factors were justified by the statistics of a validation set and can be updated on-demand.

$$S_{i,j}^{\text{local}} = \frac{1}{36}\sum_{r=i}^{i+6}\sum_{s=j}^{j+6}\left\{W_{\text{cellular\_count}} \times \frac{C_{r,s}}{5009} + W_{\text{prof.}} \times P_{r,s}^{\text{prof.}} + W_{\text{mitosis}} \times \frac{\sum \mathbb{1}[P_{r,s}^{\text{mitosis}} \geq 0.66]}{4}\right\} \tag{3.2}$$

$$S_{i,j}^{\text{HPF}} = W_{\text{cellular\_count}} \times \frac{C_{i,j}}{5009} + W_{\text{prof.}} \times P_{i,j}^{\text{prof.}} + W_{\text{mitosis}} \times \frac{\sum \mathbb{1}[P_{i,j}^{\text{mitosis}} \geq \theta_{\text{Sensitivity}}]}{4} \tag{3.3}$$

$$S_{\text{cell}} = \mathbb{1}[P_{i,j}^{\text{mitosis}} \geq \theta_{\text{Sensitivity}}] \times [P_{mitosis} - 3 \times U_{\text{mitosis}}] \tag{3.4}$$

3.6 Technical Evaluation

We conducted a technical validation study and reported the performance of the three AI models in NaviPath. Specifically, we applied classification models for mitosis and proliferation probability on the eight test scans selected from [12]. We cross-referenced the AI results and ground-truth labels to calculate F1 scores. The ground-truth labels for mitosis detection and proliferation probability calculation were acquired/generated from the annotations provided in [12]. For the cellular count calculation, we applied the model to 50 randomly-picked areas (size=512×512 pixels under $\times$400 magnification) from pathology scans. Then we compared the AI result with the cellular count reported by a graduate student, who had been briefly instructed by a pathologist (experience = 10 years).

The results showed that the mitosis detection model achieved an F1 score of 0.673 (precision: 0.703, recall: 0.650) when using a probability threshold of 0.85. The F1 score for

the proliferation probability model was 0.472 (precision: 0.544, recall: 0.416, probability threshold: 0.77). The average error rate of the cell counting model was 14.95%.

Although we tried to train the model for mitosis detection following a recent work [88], the performance of the mitosis AI was still not perfect: tuning down the threshold and setting the recall as 0.85 caused the precision score to drop to 0.216. That is, the number of false-positive instances would have been 3.62× the true-positive ones. The proliferation probability model performance was also not satisfactory, likely due to the misalignment in label distribution between train/validation and test sets: while 15.0% of train/validation data were positive, only 4.7% of test data were positive.

3.7 Work Sessions with Pathologists

We conducted work sessions with medical professionals in pathology to validate NaviPath, studying three research questions:

- **RQ1**: Can NaviPath (as a human + AI approach) increase pathologists' precision and recall in identifying the pathological features (in this case, mitosis)?
- **RQ2**: Can NaviPath save pathologists time and effort?
- **RQ3**: Compared to manual navigation, what is the benefit of using NaviPath?

We designed three testing conditions to support the system validation on the three **RQ**s:

- **C1 (Human Only)**: Participants navigate a pathology scan viewer without any AI assistance;
- **C2 (Human + AI)**: Participants navigate the pathology scan with NaviPath;
- **C3 (AI Only)**: AI-automatic reporting without humans;

3.7.1 Participants

We recruited 15 medical professionals in pathology from five medical centers across two countries, including 13 residents, one fellow (P7), and one attending (P15). The participants were recruited through flyers sent in mailing lists and word-of-mouth. The demographic information of the participants is shown in Table 3.2. All participants had received at least two years of pathology residency training to be qualified for the study (average experience μ=3.47 years, Std=0.88 years). 14/15 participants had experience in seeing pathology scans before the study (daily: 3, weekly: 6, bi-weekly: 3, monthly: 1, within one year:1). The primary purpose for using pathology scans was for learning, and the most mentioned digital pathology interface was Aperio Imagescope [110].

3.7.2 Data and Apparatus

We collected eight pathology scans of canine mammary carcinoma from a public dataset [12]. The average size of these scans was 7.15 giga-pixels. We acquired the ground-truth mitosis annotations from the same dataset [12]. Overall, the average **mitotic rate** (*i.e.*, **MR**, mitotic count per unit area[5]) was 1.022/mm^2 (0.164/HPF). We selected two scans for tutorial purposes, leaving the other six for testing (Scan 1-6 in Table 3.2). To generate AI detections, the scans were pre-processed with a local server with a 24-core CPU, 64 GB memory, and an Nvidia RTX-3090 graphics card. After that, we loaded the pre-processed results into NaviPath (**C2**). For a comparison, we developed a baseline pathology scan viewer with a basic O+D design, where pathologists were required to navigate manually to evaluate mitosis activity (**C1**). During the study, we referred to the manual baseline system as '**system 1**' and NaviPath as '**system 2**' to avoid bias.

[5] `https://www.cancer.gov/publications/dictionaries/cancer-terms/def/mitotic-rate`

Table 3.2: Demographic information and arrangements of the participants in the work sessions. The number '1' indicates that the scan was examined with system 1 (baseline manual system), while '2' was with system 2 (NaviPath). MC1-3 are located in one country, and MC4-5 are in another.

ID	Years of Experience	Frequency of Seeing Pathology Scans	Medical Center	Scan 1	Scan 2	Scan 3	Scan 4	Scan 5	Scan 6
P1	4	Weekly	MC1	2				1	
P2	3	Never	MC2	1				2	
P3	4	Bi-Weekly	MC3	2				1	
P4	4	Weekly	MC3	1				2	
P5	3	Daily	MC4				2		1
P6	2	Weekly	MC1			1	2		
P7	5	Daily	MC3			1	2		
P8	4	Bi-Weekly	MC3			2	1		
P9	4	Daily	MC3		1				2
P10	3	Weekly	MC4		1				2
P11	2	Bi-Weekly	MC4		2				1
P12	3	Weekly	MC4		2				1
P13	3	Monthly	MC4			1			2
P14	3	Within One Year	MC4			2	1		
P15	5	Weekly	MC5		2	1			

3.7.3 Task and procedure

All sessions were conducted online over Zoom. Participants were first shown a tutorial video (∼10 minutes) of the manual baseline system and NaviPath. After they had watched the video, they were given links to both systems, which were accessed through the web browser. Next, each participant was instructed to perform a pathology task of assessing the mitotic activity of one pathology scan using system 1/system 2, and another with system 2/system 1. During the formative study, we discovered that pathologists might memorize the hot-spot areas of a pathology scan that they had examined before by recognizing tumor contours, even after several months. Therefore, instead of letting a participant see the same scan after a wash-out period, we instructed participants to read different scans in the work sessions (see Table 3.2). The order of seeing the scans in each session was counterbalanced across participants. During each session, participants were required to evaluate the mitotic activity following the College of American Pathologists (CAP) cancer protocol[6], which is similar to how pathologists examine the scan in practice. Finally, participants entered a post-study structured interview that included a set of Likert questions and short answers. The average duration of each study was about 65 minutes.

3.7.4 Measurements

We collected three sources of responses from users during the work session: first, we recorded participants' interactions with both systems. Second, after they had finished examining each scan, we saved participants' reportings of mitoses. Third, from the final interview, we collected participants' responses to the questionnaire. Following previous HCI research on pathology navigation [174] and pathology AI [43], we investigated the research questions with the following measurements:

For **RQ1**, we obtained the participants' mitosis reportings with the baseline **C1**, NaviPath

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

(**C2**), and AI (**C3**). We then cross-referenced them with ground-truth mitosis labels and calculated precision and recall scores. Because each participant may visit different ROIs in each trial, we individually calculated the AI's precision and recall scores (**C3**) within the areas visited by each participant in **C2**. Therefore, we can study whether the improvements in **C2** are brought by NaviPath's AI or its human-AI workflow.

For **RQ2**, we first calculated participants' average time cost on each scan. We also evaluated each participant's navigation efficiencies by counting the number of *ground truth mitosis* within the areas visited by participants in each trial and divided it by the time length. After that, we averaged the results across the participants for **C1** and **C2** individually. Here, we did not count the *mitosis reported by participants* as in **RQ1** to rule out the difference in participants' capabilities in locating mitoses. Finally, to evaluate the cognitive workload of using both systems, we asked the participants to answer two seven-scaled Likert NASA TLX questions (*i.e.*, mental demand and frustration dimensions, Table 3.3 Q1, Q2)) [100].

For **RQ3**, we first analyzed the interaction logs and summarized participants' interaction frequencies with both systems (*i.e.*, zoom, pan, selecting recommendations). What's more, we inquired about participants' ratings on system's capabilities for mitosis searching (Table 3.3 Q3), their confidence in the mitosis reportings (Table 3.3 Q4), attitudes toward using the system in the future (Table 3.3 Q5), and overall preference of system 1 *vs.*system 2 (Table 3.3 Q6).

Last but not least, to figure out whether each NaviPath component is useful for pathologists, we asked the participants to rate each component (Figure 3.8) with a seven-scaled Likert question: *(i) "Is this feature useful to your examination?"* (1= Not useful at all $\rightarrow$ 7=Very useful); *(ii) "Compared to System 1, does this feature require extra effort?"*(1=No effort at all $\rightarrow$ 7=A lot of effort).

3.8 Result and Findings

In this section, we first answer our initial research questions based on the information collected from work sessions. We then summarize the qualitative findings on pathologists' navigation traces.

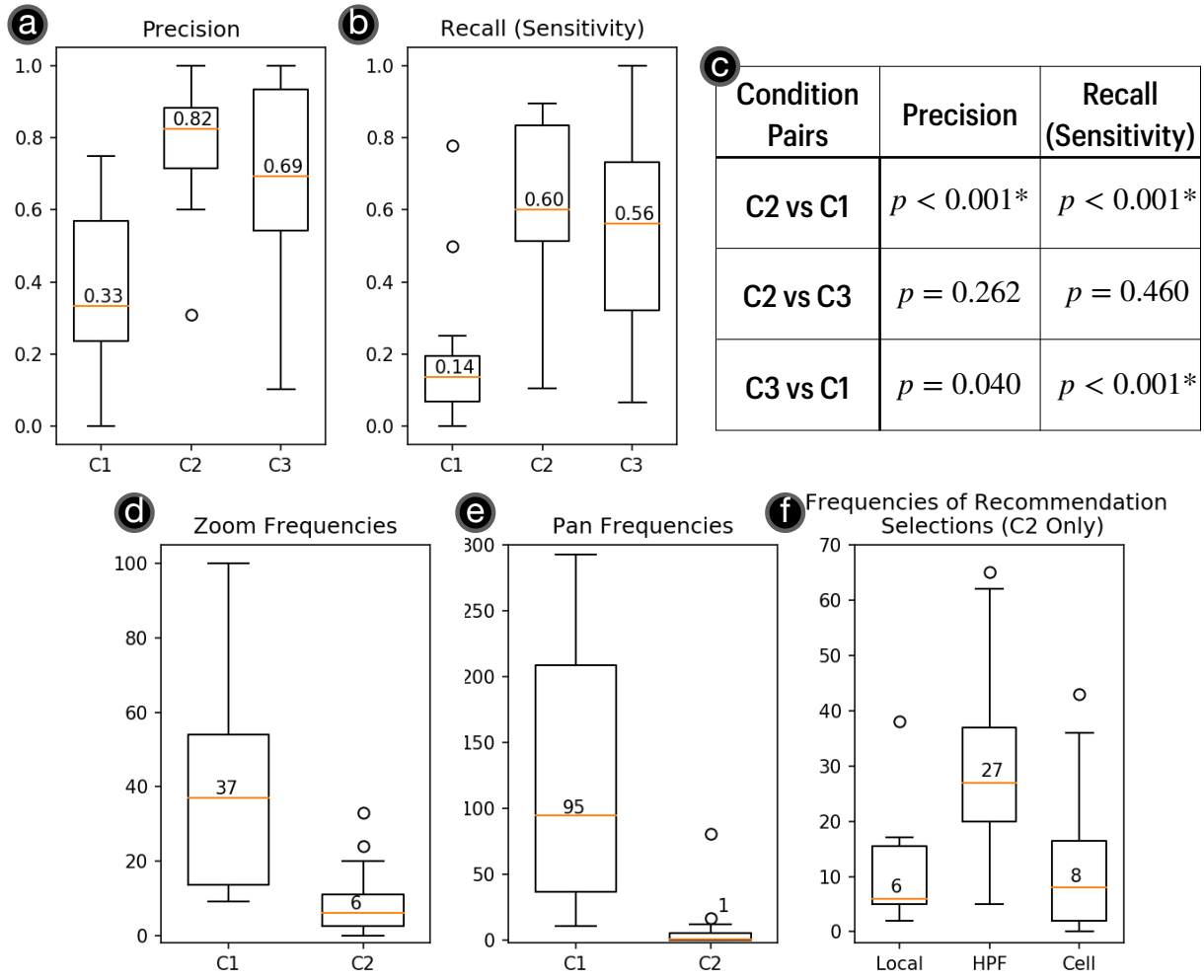
3.8.1 Results for Research Questions

3.8.1.1 RQ1: Can NaviPath increase pathologists' precision and recall in identifying the pathological features?

We calculated the precision and recall (sensitivity) of participants' mitosis reportings with manual navigation (**C1**), NaviPath (**C2**), and AI-automated reportings (**C3**) (Figure 3.7(a)-(b)). The median precision under **C1**, **C2**, and **C3** were 0.33, 0.82, and 0.69, respectively (average μ=0.40, 0.78, 0.64, standard deviation Std=0.22, 0.17, 0.31). And the median recall under the three conditions was 0.14, 0.60, and 0.56, respectively (μ=0.18, 0.61, 0.51, Std=0.19, 0.24, 0.28). An initial Kruskal-Wallis H-test indicates that precision and recall under the three conditions were significantly different (precision: p=0.002, effect size η^2_H=0.407, recall: p <0.001, η^2_H=0.511[7]). A post-hoc Dunn's test with Bonferroni correction (α=0.05) showed that recall was improved significantly when comparing **C3** *vs.***C1** and **C2** *vs.***C1** (Figure 3.7(c)). As for precision, **C2** was significantly higher than **C1**, while there was no sufficient proof to observe **C3** was higher than **C1**. We further analyzed the difference between **C2** and **C3**. On average, pathologists achieved 20.21% higher recall and 21.51% higher precision with NaviPath than AI. However, there was no sufficient proof to observe that the precision and recall were significantly higher in **C2** compared to **C3**.

It is noteworthy that participants' recall in identifying mitoses using the manual navigation is low. Upon further analysis of navigation traces, we found that the average mitotic

[7]The effect size of Kruskal-Wallis H-test η^2_H was calculated according to [196].



Condition Pairs	Precision	Recall (Sensitivity)
C2 vs C1	$p < 0.001$*	$p < 0.001$*
C2 vs C3	$p = 0.262$	$p = 0.460$
C3 vs C1	$p = 0.040$	$p < 0.001$*

Figure 3.7: Boxplot visualizations of the (a) precision and (b) recall (sensitivity) from mitosis reportings under the conditions of **C1**, **C2**, and **C3**. The colored lines and the figures above indicate the median values of each condition. The dots are the outliers. (c) The results of pair-wise significance comparison among **C1**, **C2**, and **C3** using a post-hoc Dunn's test with Bonferroni correction (α=0.05). The values marked with * indicates that the Null hypothesis can be rejected because the $p < \alpha/2$. (d) Participants' zoom interaction frequencies under **C1** and **C2**. (e) Participants' pan interaction frequencies under **C1** and **C2**; (c) Frequencies of participants' selecting Local, HPF, and Cell recommendations under **C2**. Note that one participant might select the same recommendation multiple times in each trial.

rate in the areas participants visited with the manual navigation was 0.167/HPF (which is comparable to the average mitotic rate). As a comparison, the average mitotic rate with NaviPath was 1.196/HPF, which is 6.17× higher. We believe such a significant increase (p <0.001, r=0.851, Wilcoxon rank-sum test) in the prevalence rate of the target (*i.e.*, mitosis) is the main factor why NaviPath could increase participants' recall: as described in [216], the low target prevalence would cause shifts of decision criteria that lead humans to miss targets in the visual search. NaviPath harnesses AI to recommend highly-mitotic areas for users, which brings up the prevalence rate of the visual search targets, thus helping participants achieve higher recalls (even compared with AI).

High variances in precision and recall were observed when comparing **C2** and **C3**. We believe this was caused by two factors: *(i)* variation in user interaction: in **C2**, participants chose a different recommendation customize settings and select a different amount of recommended ROIs in each trial (Figure 3.7(f)-HPF). Variations in users interactions may also result in high variance in **C3** because the precision/recall in **C3** was calculated within the areas that participants visited in **C2**; *(ii)* Variation in user's experience: different participants might adapt different thresholds to call a cell as positive.

To conclude, NaviPath achieved significantly higher precision and recall in identifying mitoses compared to manual navigation. Moreover, NaviPath, as a human + AI approach, might bring improvements compared to the AI-only condition: NaviPath achieved higher precision and recall on average. However, we did not observe that such an improvement was statistically significant.

3.8.1.2 RQ2: Can NaviPath save pathologists' time and effort?

On average, participants spent 10min27s in each trial with the baseline system, and 13min8s with NaviPath. A Wilcoxon rank-sum test indicated no sufficient proof to conclude that

Table 3.3: Summary of participants' questionnaire responses for the baseline and NaviPath with seven-scaled Likert questions. p indicates the p-value of Wilcoxon test, and r stands for the effect size. The numbers on the right indicate the averaged scores with their standard deviations. For Q1 – Q5, 1=Not at all ... 4=Neutral, ... 7=Very. For Q6, 1=Very strongly prefer system 1 over system 2, 2=Strongly prefer system 1 over system 2, 3=Slightly prefer system 1 over system 2, ... 4=Neutral, ..., 7=Very strongly prefer system 2 over system 1.

ID	**Question**	**Baseline**	NaviPath	p	r
Q1	How hard did you have to work mentally to accomplish the tasks?	5.13(1.30)	2.93(1.10)	< 0.001	0.658
Q2	How would you describe your frustrations during the tasks?	4.07(1.91)	2.40(1.06)	0.024	0.412
Q3	How capable is the system at helping count mitosis?	2.79(1.63)	6.43(0.65)	< 0.001	0.704
Q4	How confident do you feel about your accuracy?	4.21(1.42)	5.93(0.73)	0.004	0.530
Q5	Would you like to use the system in the future?	4.13(1.92)	6.47(0.64)	0.001	0.594
Q6	Overall Preference	6.33(0.82)		N/A	

participants' examinations were significantly longer (p=0.09, effect size r=0.306[8], Wilcoxon rank-sum test, same following). We further calculated each participant's navigation efficiency. The results showed that participants saw significantly more mitoses in unit time with NaviPath

compared to manual navigation (manual: μ=0.012 mitoses/second,

NaviPath: μ=0.028 mitoses/second, p=0.002, r=0.579). Specifically, NaviPath's Local recommendations served as a shortcut that guided participants directly to highly-mitotic areas without manual searching: *"The local recommendations have more mitosis inside, and I can focus on this area. I can start counting from there and I do not need to find one myself."*(P1) *"It (*NaviPath*) tells you which ones are the highest areas. And then you just go from there and decide. With system 1, you still have to review the whole slide."*(P3)

[8]The effect size of the Wilcoxon Test r is calculated as $r = \frac{Z}{\sqrt{N}}$, where Z is z-score from the Wilcoxon Test, and N is the number of observations (30 in this study).

In the post-study questionnaire, participants reported significantly less mental effort with NaviPath (manual: μ =5.13, NaviPath: μ =2.93, p <0.001, r=0.658) compared to the manual navigation (Table 3.3 Q1). Furthermore, participants expressed less frustration using NaviPath (manual: μ =4.07, NaviPath: μ =2.40, p=0.024, r=0.412, Table 3.3 Q2). Specifically, participants valued NaviPath's Cell recommendations as the key to reducing the workload — *"It (*NaviPath*) takes away the burden of seeing and hunting for mitosis... it can tell you where is most likely to have mitosis and you decide 'yes' or 'no'."*(P3)

In sum, although participants spent longer time using NaviPath on average, their navigation efficiency was improved significantly by NaviPath's Local recommendations — they could see more than twice the number of mitosis in unit time. Moreover, according to the questionnaire response, participants reported significantly less effort when using NaviPath. NaviPath's Cell recommendations contribute the main improvement: they could highlight specific cells from a large background, freeing pathologists from tedious manual visual search.

3.8.1.3 RQ3: Compared to manual navigation, what is the benefit of using NaviPath?

We answer this question by first comparing the patterns of interactions (*e.g.*, pan, zoom) while participants use NaviPath (**C2**) *vs.*with the manual navigation (**C1**). In sum, zooming and panning made up most of participants' interactions under **C1**, while "selecting AI recommendations" took the majority of interactions under **C2** (NaviPath). The median frequencies of zoom interactions under **C1** and **C2** were 37 and 6 (Figure 3.7(d)). And the median pan interaction frequencies under **C1** and **C2** were 95 and 1 (Figure 3.7(e)). A Wilcoxon test showed that zoom and pan interactions were significantly reduced under **C2** (zoom:p <0.001, r=0.651; pan: p <0.001, r=0.784). Furthermore, with NaviPath, participants selected a median of 6 Local, 27 HPF, and 8 Cell recommendations in each trial.

According to the questionnaire responses, participants believed that NaviPath was more

capable of assisting in detecting mitosis (manual: μ=2.79, NaviPath:μ=6.43, p <0.001, r=0.704, Table 3.3 Q3). Pathologists' confidence in mitosis reportings was improved significantly by NaviPath (manual: μ=4.21, NaviPath:μ=5.93, p=0.004, r=0.530, Table 3.3 Q4). Specifically, participants expressed that the AI recommendations would serve as a second opinion while they made justifications — *"I was kind of like 90% sure ... but then if AI was 100% sure, I felt more confident in saying that it was real mitoses."*(P3). *"It's kind of like having a second set of brains."*(P6). Finally, participants expressed that they were more likely to use NaviPath in the future (manual: μ=4.13, NaviPath:μ=6.47, p=0.001, r=0.594, Table 3.3 Q5). Overall, as shown in Table 3.3 Q6, participants indicated a preference for system 2 (NaviPath) over system 1 (baseline pathology scan viewer): based on the questionnaire, 8/15 of the participants rated a score 7 (very strongly prefer system 2 over system 1), 4/15 rated a score 6 (strongly prefer system 2 over system 1), and 3/15 rated a score 5 (slightly preferred system 2 over system 1).

In sum, users could navigate the pathology scans by selecting AI recommendations from NaviPath. Meanwhile, their pan and zoom interactions were significantly reduced. Overall, they believed NaviPath was more capable of finding mitosis, had higher confidence while using NaviPath, and preferred to use it in the future.

3.8.2 Ratings on NaviPath's Components

To further understand whether each NaviPath component was useful for pathologists, we asked participants to rate each (see Figure 3.8). Here, we report the participants' ratings and discuss qualitative findings, organized by the categories of components:

3.8.2.1 Hierarchical AI Recommendations

Participants rated average useful ratings of 5.93/7, 6.53/7, and 6.53/7 for Local, HPF, and Cell recommendations, respectively. Specifically, participants expressed that Local and HPF

Category	Items	Is this feature useful to your examination? (1: not useful at all → 7: very useful)									Compared to system 1, does this feature require extra effort? (1: not effort at all → 7: a lot of effort)								
		1	2	3	4	5	6	7	Mean	Std	1	2	3	4	5	6	7	Mean	Std
Hierarchical AI Recommendations	Local			2	1	1	3	8	5.93	1.49	8	5	1	1				1.67	0.90
	HPF					1	5	9	6.53	0.64	5	8	1	1				1.87	0.83
	Cell				1		4	10	6.53	0.83	5	8		1	1			2.00	0.13
Customizable Recommendation by Multiple Criteria	Cellular Count			2	2	3	3	5	5.47	1.46	7	5	1	2				1.87	1.06
	Proliferation Probability			1	2	3	3	6	5.73	1.33	7	5	1	2				1.87	1.06
	Mitosis Count			1	1	2	5	6	5.93	1.22	7	5		3				1.93	1.16
	Mitosis Sensitivity				2	2	5	6	6.00	1.07	7	5		2	1			2.00	1.49
Cue-Based Navigation	Navigation Cue		1		2	8	4		4.93	1.39	6	6	2	1				1.87	0.92
Explanation for Each Recommendation	Verbal Dialog			2	4	2	4	3	5.13	1.41	5	5		4	1			2.40	1.40
	Explanation Card				2	2	7	4	5.87	0.99	3	5	2		4	1		3.00	1.73

Figure 3.8: Participants' ratings on whether each component in NaviPath is useful to pathologists' examination (left) / requires extra effort compared to the manual baseline system (system 1) (right).

recommendations helped them narrow down from a large region without manual navigation — *"The entire slide might have thousands of high-power fields, and the Local recommendations picked the highest 36 for me ... the HPF recommendations continued to pick about 20 high-power fields from the Local recommendation ... it helps me rule out regions and focus on the important areas."*(P14)

Notably, Cell recommendations received the highest useful rating among NaviPath's components. Participants expressed that Cell recommendations transformed the task of visual search into adjudication, which can save their mental effort. Specifically, they used Cell recommendations as an additional layer to quickly locate and adjudicate suspected cells: for most scenarios, participants directly reported the mitosis after glancing at the Cell recommendations. If they were not confident, they continued to select a Cell recommendation and examine it closely with a higher magnification. This explains why Cell recommendations were rated most useful, although they were not selected frequently in practice (as reported in Section 3.8.1.3).

3.8.2.2 Recommendation Customization by Multiple Criteria

Amongst the three criteria that NaviPath used to generate recommendations, participants gave the "mitosis count" the highest usefulness rating (μ=5.93/7), followed by the "proliferation probability" (μ=5.73/7) and "cellular count" (μ=5.47/7). Although most participants expressed that all three criteria should be considered in general, some (P2, P4, P15) believed it was not challenging for human pathologists to pick cellular areas, and it was not highly motivated to employ AI as such.

We also found that participants did not frequently interact with the slide-bars to change the recommendation customization settings for the three criteria. Instead, they picked a custom set-up at the beginning of each trial and left them unchanged. Upon further analysis, we found that NaviPath's recommendations might not change after users moved the slide-bars under certain circumstances, which disincentives users' interactions — *I don't see it (the recommendation) changing much when I set the 'cellular count' as 'high'."*(P1) What's more, adjusting the customization settings during the examination might incur extra workload, and P14 suggested NaviPath give pre-set values for the three criteria — *"It would be great if the system could give me default values for the three criteria ... changing the criteria is a lot of work if I see hundreds of slides."*

Furthermore, participants had diverse opinions on how much a criterion should be considered in AI recommendations. One participant only gave "mitosis count" a high weight while giving zero weight for the other two criteria: *"I want AI to go straight to the mitoses, not like just predict for me based on the cell count where there are more mitoses elsewhere."*(P4) However, others thought NaviPath should also include other criteria for recommendations. For example, P6 gave both "cellular count" and "mitosis count" a high weight — *"I would like to include the cellular counts ... this is how we see tumors every day."*(P6)

As for the sensitivity slide-bar, participants usually set it as "high" to see more recommendations, although this may produce false positives: *"I move it all the way to the right,*

it will detect more mitosis ... not all of them will be real mitosis, but it has more sensitivity. So then I can decide if the real to me or not."(P3) Pathologists' preferences of recall (sensitivity) over precision was also reported in Chapter 4. We believe such preferences are rooted from the imbalance risks in pathology decision making: while a proliferation of false-positive results (from low threshold) may cause longer time in examination, false-negative results (due to using a high threshold) might make the diagnosis unreliable because of the failure to acknowledge critical pathological features.

3.8.2.3 Cue-Based Navigation

Surprisingly, the navigation cue received the lowest usefulness ratings by participants, with an average score of 4.93/7. Participants' opinions were split into two groups when asked how they used the navigation cue during work sessions. On one hand, some participants (P5, P10, P14) used cue-based navigation during their examination, and treated the navigation cue as a short-cut to access possible mitosis areas — *"It allows me to quickly locate the area where the next possible (mitosis) is located."*(P5). On the other hand, some participants expressed that the cue-based navigation might be incompatible with a medical guideline: *"I sometimes did not know where these cues would guide me to ... because we need to see (mitoses in) 10 consecutive areas. And I didn't know if I was jumping from one to the other at the end they wouldn't be really consecutive"* (P1) Regarding how participants might navigate under the high magnifications with NaviPath, we will discuss in more detail in Section 3.8.3.2.

3.8.2.4 Explanations for Recommendations

Participants gave average ratings of 5.13/7 in usefulness and 2.40/7 in effort for the verbal explanation dialog. P5, P6, P7, P11, and P12 expressed that the verbal dialog assisted them in prioritizing the examination of HPF recommendations — *"Here (pointing at one HPF recommendation), it (the verbal dialog) says 'very cellular' and 'moderately likely'.*

And then here (pointing at another HPF recommendation), it says 'very cellular' and 'very likely'. So I might pick this box (the latter one) to see first ... it will be helpful to my selection."(P6) However, four participants (P10, P13, P14, P15) ignored the verbal dialog during the examination and used the ranking indexes to select HPF recommendations instead — *"I think the verbal dialog and the recommendation rankings are redundant ... the rule says the lower the (ranking) number, and more important the box is ... I feel that the ranking numbers are more straightforward."*(P15)

As for the explanation card, participants gave a usefulness rating of 5.87/7. If participants were not confident about whether a Cell recommendation was mitosis, they would refer to the explanation card as a confirmation: *"I just took it as confirmatory that my assessment was correct."* (P8) It is noteworthy that the explanation card also received the highest effort score (3.00/7) among NaviPath's components because participants spent extra effort comprehending the explanations.

3.8.3 Qualitative Findings on Participants' Navigation Traces

We analyzed participants' navigation traces on the pathology scans and report the qualitative findings on pathologists' navigation traces with the manual baseline system and NaviPath.

3.8.3.1 Navigating the scan manually *vs.*with NaviPath

One notorious issue of the pathology examination is the low between-subject consistency, which is usually caused by the randomness in pathologists' navigation. We also observed such randomness during our user study. For example, Figure 3.9(a) visualizes the 2D projections of three (P5, P11, P12) participants' navigation traces with the manual navigation. It is noteworthy that all three traces barely overlap, which might result in inconsistencies in the medical decision makings. Also, all three participants did not examine a tissue session on the bottom-right corner of the scan (pointed by the arrow). However, according to the

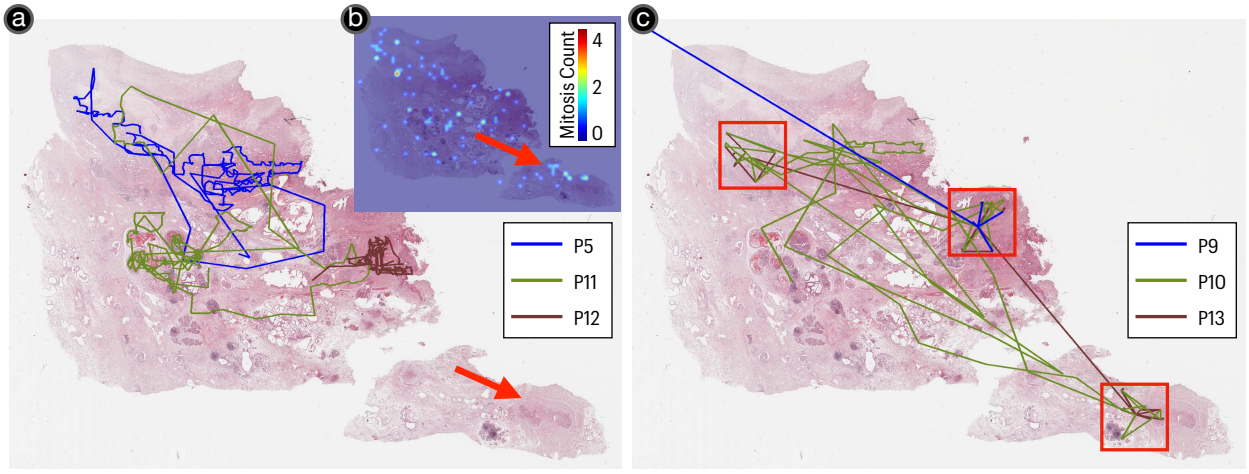


Figure 3.9: 2D projections of participants' traces with manual and NaviPath navigation on a pathology scan (zoom ignored). (a) Trace projections of P5, P11, and P12 with manual navigation. Note that all three participants did not examine the tissue on the bottom-right corner of the scan (pointed by the arrow). (b) The heatmap visualization of mitosis density of the scan. (c) Trace projections of P9, P10, and P13 with the NaviPath navigation. The boxes highlight the approximate areas of Local recommendations generated by NaviPath.

ground-truth mitosis density heatmap (Figure 3.9(b)), the unexamined tissue session has aggregations of mitoses (shown as hotspots, pointed by the arrow). Therefore, the decisions made with the manual navigation might be biased because one important area was missed.

In contrast, participants' traces are more consistent with NaviPath. Figure 3.9(c) illustrates three other participants' navigation traces (P9, P10, P13) within the same scan with NaviPath navigation. The boxes indicate the approximate areas of Local recommendations generated by NaviPath. Thanks to AI recommendations, participants' navigation traces are more consistent within the three Local recommendations. Also, P10 and P13 examined the tissue session that had been missed in the manual navigation.

Therefore, NaviPath can improve participants' consistency and also increase the exploration of their navigation.

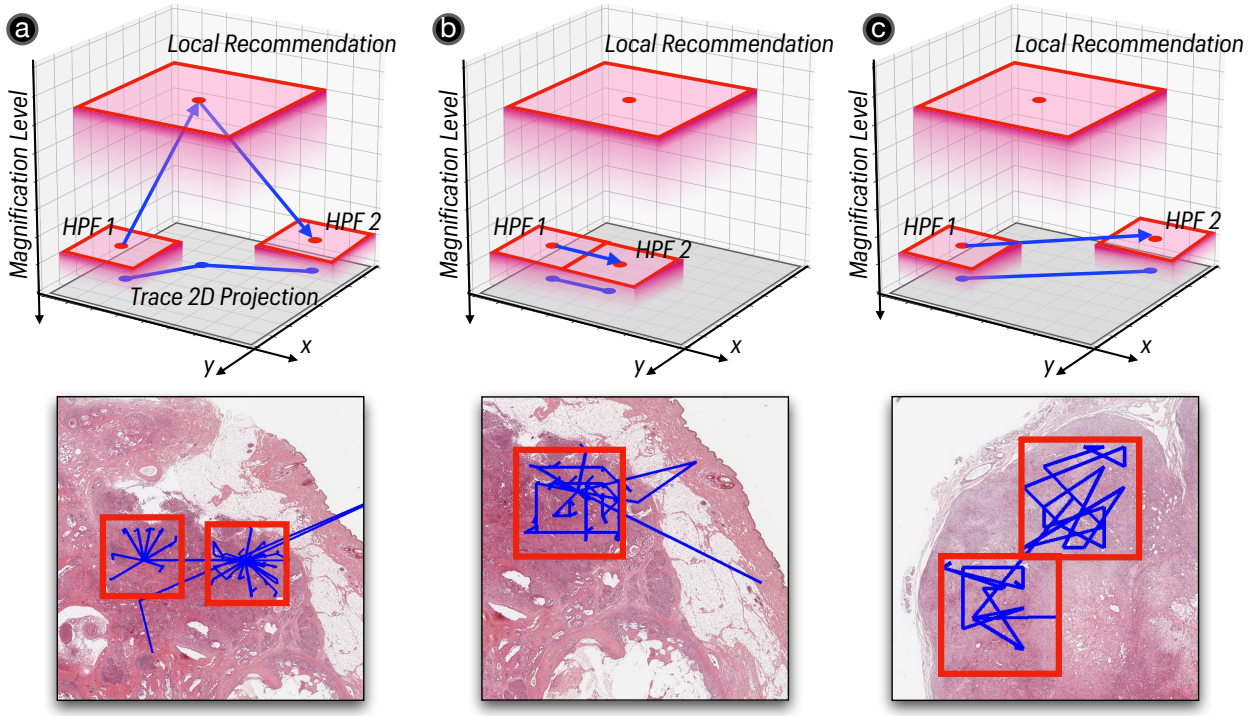


Figure 3.10: Three patterns of how our participants move to another HPF recommendation after examining one: (a) "Diving": first returned to the Local recommendation, overviewed the remaining HPF recommendations from the low magnification, and then dived down by selecting an HPF recommendation. The bottom figure shows 2D projections of participants' navigation traces during the work sessions; (b) "Adjacent Panning": directly pan to an adjacent HPF recommendation by clicking on the edge of NaviPath's interface; (c) "Cue-Based Hopping": directly hop to a remote HPF recommendation with the navigation cue.

3.8.3.2 Moving from one HPF recommendation to another with NaviPath.

From the formative study, we learned that pathologists searched systematically in high magnifications with manual navigation. Here, we study whether our participants' navigation patterns in high magnifications with NaviPath are different: specifically, we analyzed participants' navigation traces and summarized three navigation patterns of how our participants moved to another HPF recommendation after examining one:

- **Diving**: Participants first moved to the Local recommendation, then overviewed remaining HPF recommendations with low magnification, and selected an HPF recommendation to examine in higher magnification (Figure 3.10(a)). During work sessions, P8 and P15 mainly used the diving navigation, and would switch the magnifications by selecting NaviPath's hierarchical recommendations without getting lost. As shown in Figure 3.10(a), the bottom figure, the diving navigation left a 'spoke-like' navigation trace (the blue line) within each Local recommendation (red boxes).

- **Adjacent Panning**: Participants clicked on the edge of NaviPath's interface to move discretely to an adjacent HPF recommendation (Figure 3.10(b)). The adjacent panning is the closest to current pathologists' navigation practices (without AI), and five participants (P2, P3, P4, P7, P11) employed the adjacent panning in the study. The navigation trace is more regular with the adjacent panning (see Figure 3.10(b), the bottom figure).

- **Cue-Based Hopping**: Participants clicked on the navigation cue to hop to a remote HPF recommendation (Figure 3.10c). P5, P10, and P14 mainly used it during the study. With cue-based hopping, participants were able to see the HPF recommendations in ascending order based on ranking index to maximize navigation efficiency — *"My preference is to click on the navigation cue and jump to the next important HPF. For example, after I have seen number 1 (HPF recommendation), I will see num-*

ber 2."(P10) As shown in Figure 3.10(c), the navigation trace is more irregular with cue-based hopping.

3.9 Discussion

3.9.1 Limitations

3.9.1.1 Limitations of the evaluation study

- **User Sampling**: The majority of participants are pathology residents with relatively less experience, making the conclusions for **RQ1** inevitably speculative due to a lack of participation of more-experienced attending pathologists;
- **Study Set-Up**: The work sessions were relatively brief because of the scarce availability of participants, and no clinical experiments were conducted because of strict regulations from US Food and Drug Administration (FDA);
- **Materials**: All pathology scans used in the study have the same tumor type because of the rare availability of public datasets. Therefore, they lack variability to reflect the real-world distribution of pathology data;
- **Choice of Baseline**: No comparison between NaviPath and other human-AI systems was conducted because there is a lack of open-source systems for mitosis detection. There was also no comparison conducted with the optical microscope, pathologists' primary approach to see tumor specimens, due to the COVID-19 pandemic.

Therefore, future works should concentrate on conducting larger-scaled, longer-termed, in-the-wild studies to evaluate the influence of implementing a human-AI collaborative navigation system for pathologists.

3.9.1.2 Limitations of NaviPath

- The two deep learning models for the proliferation probability and mitosis classification were trained from images of one tumor, and their performance on other tumors is unknown;
- The current cue-based navigation design used in NaviPath (*i.e.*, citylight) cannot provide the distance information of off-screen recommendations, and might be incompatible with specific medical guidelines;
- The current recommendation customization algorithm was not predictable under certain circumstances;
- NaviPath does not support users to add their own ROIs for examination. Thus, users need to examine manually if an area was not recommended.

As such, future work should train AI models from various tumors to improve the model's generalizability. And future systems might consider other cue-based navigation designs (*e.g.*, Wedge [94] or Halo [24]) that can offer both distance and directional information of off-screen targets, which can support navigation according to medical guidelines. Another improvement direction is modifying the overview map in the O+D design: by demonstrating where the pathologist is looking and all recommended ROIs to enhance humans' spatial awareness of off-screen targets (*e.g.*, [34]). Future works should also consider utilizing machine intelligence to support the examination of user-defined ROIs: for example, a user can select an area of interest manually, and the system can recommend all salient AI findings inside for the user to examine [60]. Finally, we also suggest future works to improve the predictability of medical AI, which we will discuss next.

3.9.2 Implications for Human-AI Designs in Medical Decision-Making

3.9.2.1 Making AI-Enabled Systems Predictable

Previous work suggests that the disruptive behavior of AI might discourage medical professionals from using it in practice [222]. In our study, we discovered that participants did not change the customization settings frequently because the outcomes were less predictable: for example, tuning the "Cellular Count" slide-bar would simultaneously change recommendations' locations and rankings. In some scenarios, tuning the slide-bar would not change the recommendations at all.

It is challenging for doctors to be aware of whether the change is beneficial or the no change is caused by malfunction. As such, we suggest future human-AI systems in medicine to present intuitive clues that aid doctors in evaluating changes made by AI. For instance, future systems can justify why changes are happening or not – text explanations generated by NLP agents (similar to [210]) can be implemented to explain the AI status and help pathologists comprehend the recommendation reasoning process. Another future direction might include making the recommendation AI less disruptive: for example, recommendations based on human-understandable medical concepts can make the algorithm more predictable for medical users [43].

3.9.2.2 Balancing Simplicity and Informativeness

Doctors prefer simple, straightforward designs [89]. From the evaluation study, we found that some participants preferred to use the ranking index number over the verbal explanation dialog. However, simpler designs usually mean "lossy" information compression, and might not be sufficiently informative for medical decision-making. Therefore, we suggest future HCI research to study what information should be preserved *vs.*discarded through empirical studies. For instance, in the next chapter, we will show pathology AI systems provides levels of AI explanations for doctors: a simple, visual explanation was shown by default,

while more detailed explanations could be retrieved on demand. By balancing simplicity and informativeness, doctors can rapidly inquire about the most salient information with less confusion.

3.9.2.3 Decoupling Doctors and AI

Recent research has reported that utilizing AI may cause doctors' diagnoses to align with that of AI's [76]. However, it is still unknown whether the alignment is beneficial or catastrophic because the performance of AI is subject to be influenced in clinical settings [26]. Moreover, previous research suggests that the domain gap in pathology image data will harm AI performance [187, 11]. Therefore, doctors only examining within the AI-recommended areas would put physician-AI collaboration into a dilemma: on one hand, they may miss critical findings if the model's recall (sensitivity) is less than 1.00; on the other hand, seeing all areas comprehensively can barely reduce human workload. To tackle this problem of speed and accuracy, future improvements might consider re-designing the human-AI collaborative workflow: doctors might first overview a medical image and generate an overall impression of the case, then a human-AI collaborative system can be engaged to enable doctors to verify or refine their initial hypotheses [39]. What's more, providing additional sources of information might be an improvement: for example, attaching immunohistochemistry tests along with conventional pathology scans can let pathologists justify whether AI recommendations are reliable. Another unresolved question in this chapter is, since various pathological patterns might co-exist in a scan, are pathologists required to see other pathological patterns after examining one with NaviPath? In short, it depends on whether the criterion (in this work, mitosis) is deterministic for diagnoses according to the medical standard, and we will discussion with more details in the next chapter.

3.10 Conclusion

This chapter introduces NaviPath to enhance pathologists' navigation efficiency in high-resolution tumor images by integrating domain knowledge and taking account of a practical workflow based on an empirical study with medical professionals. NaviPath could save pathologists from repetitive navigation in high-resolution tumor images through its AI-enabled designs. In contrast to prior work, we center on pathologists and adapt AI tools into their workflow to facilitate navigation processes. NaviPath mainly focuses on mitosis in pathology, which represents a class of highly challenging problems on domain-specific navigation with high-resolution images. In the next chapter, we will discuss the design and validation of AI-assisted system for more complex, multi-criteria diagnosis tasks.

CHAPTER 4

Advancing Multi-Criteria Decision Support in Pathology through Human-AI Collaboration

This chapter is based in part on the following publication:

Hongyan Gu, Yuan Liang, Yifan Xu, Christopher Kazu Williams, Shino Magaki, Negar Khanlou, Harry Vinters, Yang Li, Mohammad Haeri, and Xiang 'Anthony' Chen "Improving workflow integration with XPath: design and evaluation of a human-AI diagnosis system in pathology." *ACM Transactions on Computer-Human Interaction* 30, no. 2 (2023): 1-37.

4.1 Introduction

The past decade has experienced rapid development in digital pathology, which transforms physical glass slides into high-resolution digital whole slide images (WSIs) [157]. This transformation lays the foundation for assisting diagnoses with machine intelligence [6, 40, 96], and might improve patient management ultimately [28]. To date, AI (Artificial Intelligence) has been proposed for a broad spectrum of potential applications of pathology [211, 106, 163, 191, 212], with some achieving performance on par with human beings in labs [27, 228]. Furthermore, various AI models have been adopted into tools to support pathologists' tasks, targeting automating parts of pathologists' workflow to reduce their examination burdens [44, 113, 66]. 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 [222, 141, 195, 121].

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" [44] and "connect to existing clinical processes" [113]; it also indicates "unique difficulties" in converting human-AI interaction guidelines to tool support [222]. To this end, previous literature has explored the designs and influence of human-AI collaborative workflows for medical professionals [26, 76, 127, 210]. For pathology, numerous works have revealed the potential of human-AI collaborative systems to support doctors' exploration of one or more pathological patterns [43, 113, 60]. Extending the success of previous works, this chapter 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 [113], or assisting adjudications on areas of interest [101, 43]. 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 [228], concept attributions [43],

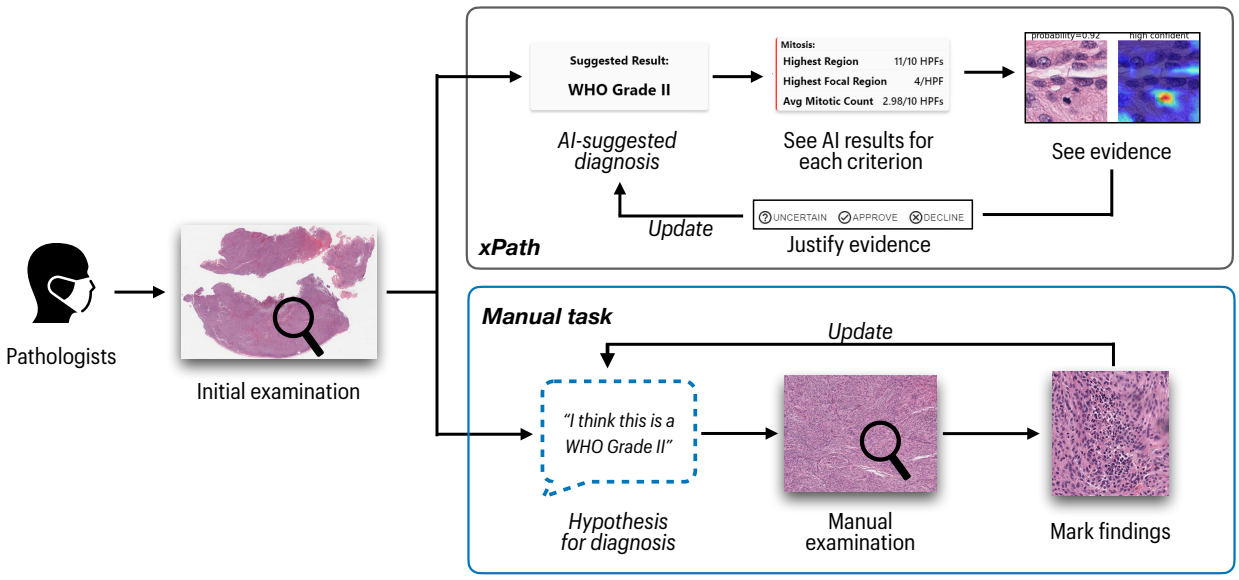


Figure 4.1: Workflow of xPath (up): pathologists first see the AI-suggested diagnosis, then examine its results and evidence accordingly in an explainable manner, and examine the evidence to update the suggested diagnosis. this workflow follows a similar manual examination process of pathologists (down), which can improve AI's integration into pathologists' routine diagnoses.

and confidence scores [72]. 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 [220], 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 [136].

Building upon the design challenges from the formative study, we propose xPath — a

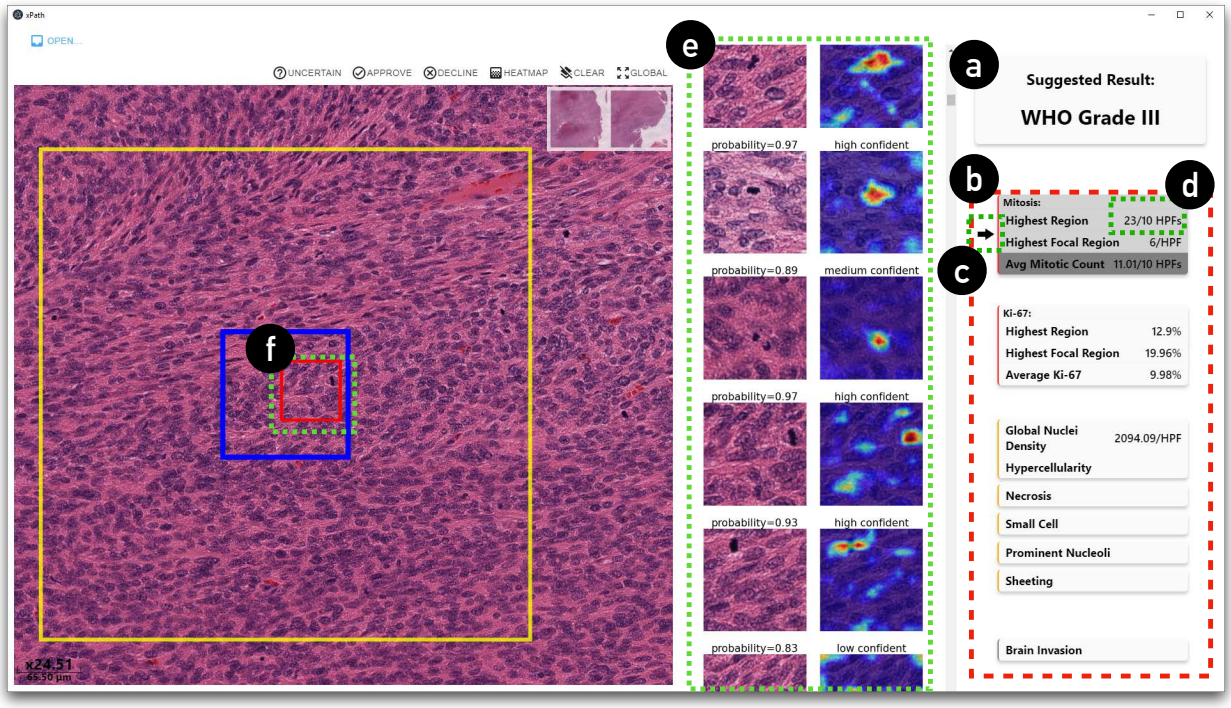


Figure 4.2: xPath's interface design, illustrating the (a) suggested pathology diagnosis (*i.e.*, WHO Grade 3) with two key design ingredients of (b) joint-analyses of multiple criteria, where xPath offers comprehensive AI analysis of multiple critical pathology criteria for a diagnosis; explanation 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.

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; *(iii)* it allows pathologists to perform diagnoses in a similar workflow to their routine practice (as shown in Figure 4.1).

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 4.2b), which are combined to produce a suggested diagnosis (Figure 4.2a) based on rules derived from the existing medical guideline [136]. 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 4.2a) with a quantified score for the criterion (Figure 4.2d), to a list of evidence that contributes to the quantified score (Figure 4.2e), and further to examine each evidence with contextual information by registering it to the whole slide image (Figure 4.2f). 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 4.1).

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 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.

4.1.1 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,

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

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.

4.2 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 4.3e) or small events (*i.e.*, mitosis, see Figure 4.3c), and, *(iii)* examine the entire tumor (occasionally as many as 20 or more slides) for frequently rare features (*i.e.*, spontaneous necrosis, see Figure 4.3i). 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 [136]. 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 [137].

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 [208]; 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 [156].

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 4.3 briefly visualizes examples of pathological features that pathologists need to find. Pathologists' work starts with the H&E slides (Figure 4.3a). Apart from the H&E, Ki-67 IHC tests [2] are often used (Figure 4.3b) to provide an estimated proliferation index (Figure 4.3d,k), which is highly correlated to meningioma grading. According to the WHO guidelines (see Section 4.2.1) [137], 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.

4.2.1 WHO Guidelines for Meningioma Grading (WHO CNS 5)

As specified by the WHO Central Nervous Tumor (CNS), 5th edition [137], meningioma grading can be based on the following criteria:

- **Grade 1** (benign) meningiomas include "histological variant other than clear cell, chordoid, papillary, and rhabdoid "[36] 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:

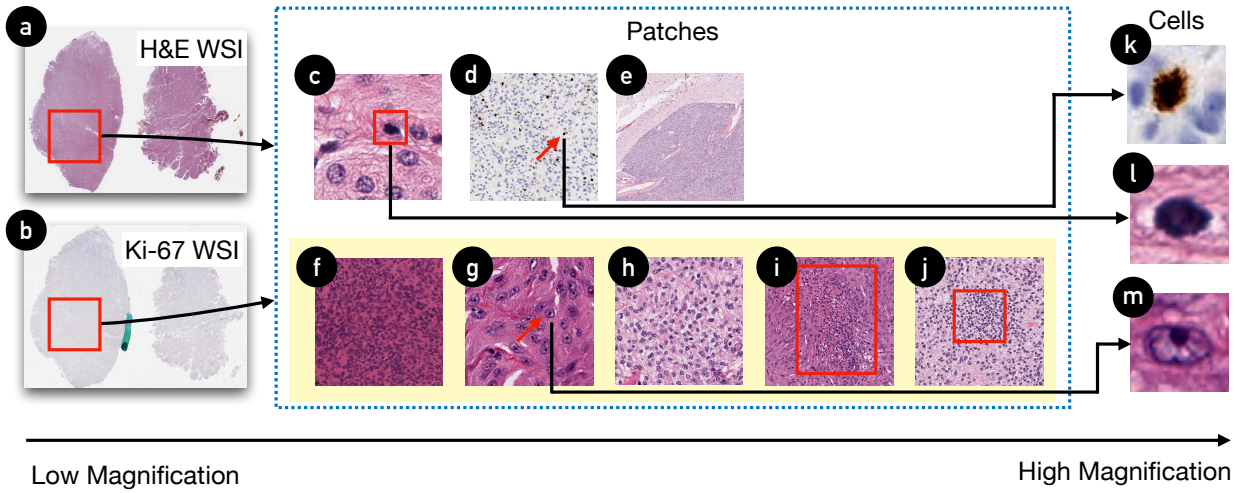


Figure 4.3: Examples of criteria used for the meningioma grading. (a) The resected tissues are first stained with H&E solution. (b) An additional Ki-67 IHC test is usually used to locate mitoses. According to the WHO grading guidelines, pathologists look for (c) mitotic cells (marked in the red box) 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.

1. The presence of ≥ 2.5 mitoses/mm^2 (equating to ≥ 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 4.3k) in the corresponding areas of Ki-67 (Figure 4.3d) 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 4.3f), prominent nucleoli — enlarged nucleoli in a cell (usually as a cluster) (Figure 4.3g,m), sheeting — loss of 'whirling' architecture (Figure 4.3h), necrosis — irreversible injury to cells (Figure 4.3i), and small cell — cluster of cells with high nuclear/cytoplasmic ratio (Figure 4.3j);
3. Brain invasion — invasive tumor cells within the brain tissue is observed (Figure 4.3e);
4. The dominant appearance of clear cell or chordoid subtype.

- **Grade 3** meningiomas are decided if at least one of the following criteria met [18, 137]:
    1. Mitotic figures of ≥ 12.5 mitoses/mm ^{2} (equal to ≥ 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.

4.3 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

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 4.1: Demographic information of the participants in the formative study.

examined meningiomas weekly. The demographic information of the participants is shown in Table 4.1. Two out of four participants (FP3, FP4) have used digital pathology systems, and the primary software they used is Imagescope[2]. 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.

4.3.1 Moderator's Questions for the Formative Study

Below is a list of questions asked by the moderator in the semi-structured interview for the formative study. Note that the order of questions and aspects discussed might vary during the study.

Part 1: Questions for asking participants' behavior for meningioma grading:

- Based on the case we have just seen, could you describe how you examined the slide

[2]`https://www.leicabiosystems.com/us/digital-pathology/manage/aperio-imagescope/`

to grade the case?

- How do you usually do to grade meningiomas?
- What problems do you think exist in meningioma grading workflow? Why do you think it is a problem?

Part 2: Questions for participants' expectations for AI-aided meningioma diagnosis system:

- Describe why you need an AI system for meningioma grading.
- Suppose there was an automated diagnosis to help you do this job, could you please give me an overview of the key functions or processes? That is, what functions/capabilities do you think is very important to the system, and why?

4.3.2 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 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)[3]. 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 [62] — "*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.

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

4.3.3 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 4.5.

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 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 hy-

pothesis 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.

4.4 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.

4.4.1 Joint-Analyses of Multiple Criteria

Based on the formative study, we found that pathologists rely on the WHO meningioma grading guideline for meningioma grading [136] involving multiple criteria. Thus xPath's design follows the WHO guideline and employs AI to compute eight critical criteria for meningioma grading[4]. Details on the AI implementation are described in Section 4.5. 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

[4]... which includes the mitotic count, Ki-67 proliferation index, hypercellularity, necrosis, small cell, prominent nucleoli, sheeting, and brain invasion. Note that this chapter 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.

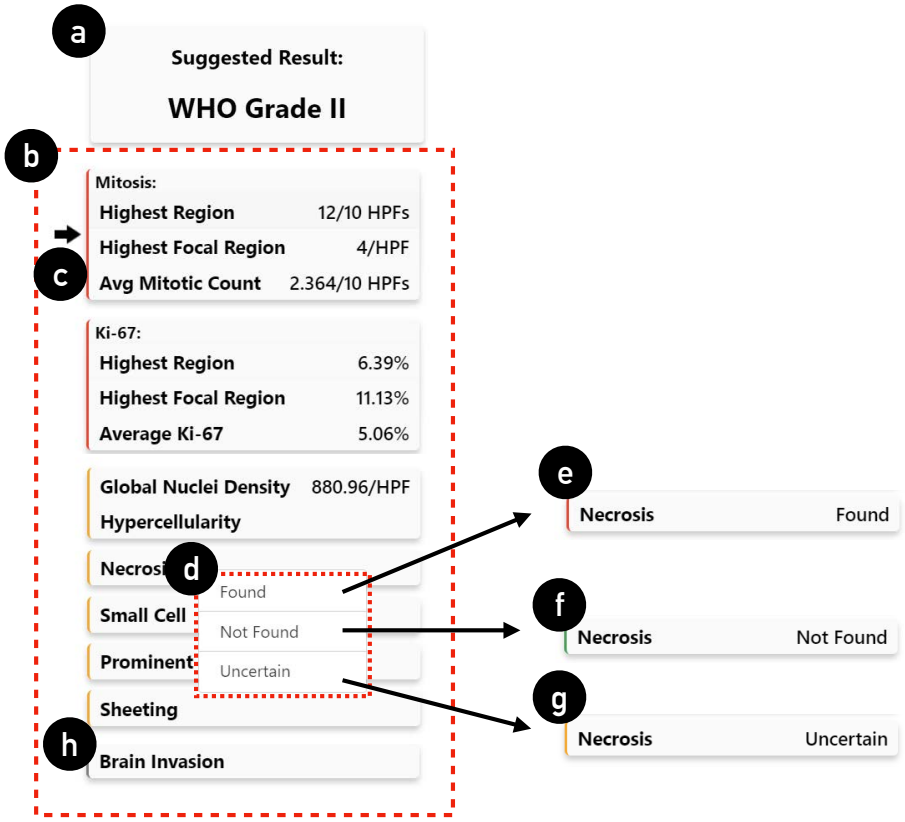


Figure 4.4: Joint-analyses of multiple criteria in xPath's design: (a) the overall suggested grading; (b) a structured overview of each WHO criterion with (c) an arrow highlighting the main contributing criterion to the suggested grading; (d) users can override criteria by right-clicking on each item 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.

according to the largest aggregations of AIs' probabilities.

Figure 4.4 demonstrates the interface of multiple criteria, which shows the current suggested grading for the tumor (*i.e.*, the suggested 'WHO grade 2', Figure 4.4a) and a structured overview of each criterion (Figure 4.4b). xPath displays an arrow to indicate the main contributing criterion (Figure 4.4c), the most deterministic AI findings for the suggested diagnosis, according to the meningioma grading guidelines. For example, in Figure 4.4, xPath suggests the "*mitotic count*" is the main contributing criterion, because it has detected 12 mitoses in 10 high-power fields (HPFs) (Figure 4.4c, 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 4.4d). And xPath uses color bars (Figure 4.4e,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 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.

4.4.2 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.

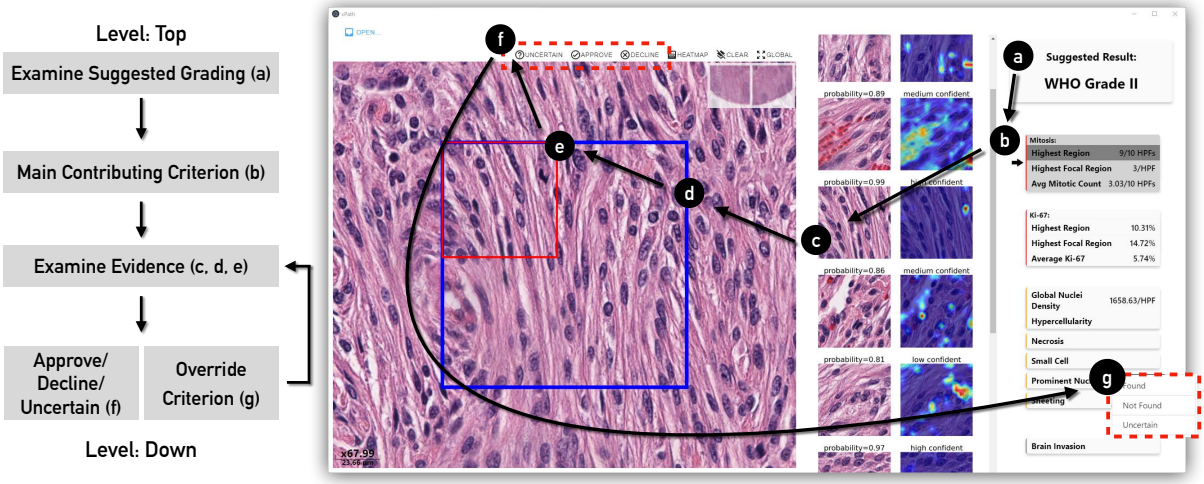


Figure 4.5: xPath presents a top-down human-AI collaboration workflow for pathologists to interact with xPath (left) and pathologists' corresponding footprints on the xPath's frontend user interface with examining the mitosis criterion as an example (right). A pathologist user starts from (a) the AI-suggested 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.

Specifically, at the **top level**, pathologists can first see the suggested diagnosis recommended by xPath (Figure 4.5a). Then, they can continue to dive down and examine a list of AI-computed criteria (Figure 4.5b). Each criterion can be boiled down to a list of **mid-level** samples (Figure 4.5c). For the most important criterion — mitosis, xPath demonstrates a series of explanations in each sample, including AI's output probability (Figure 4.6a), AI's confidence level (Figure 4.6b), and a saliency map (Figure 4.6c) that highlights the spatial support for the mitosis class in the reference image, 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 4.5d,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 4.5f), or directly override AI's results on each criterion (Figure 4.5g). Correspondingly, the overall suggested grading (Figure 4.5a) 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 [102].

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 4.7 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:

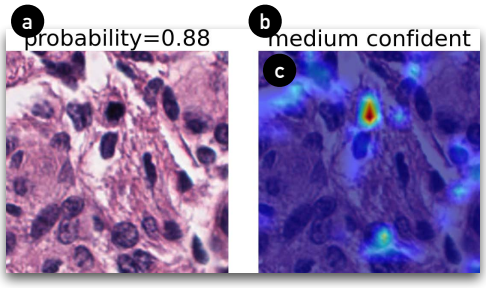


Figure 4.6: For the mitosis criterion, xPath demonstrates a series of explanations in each mid-level sample, including the (a) AI's probability, (b) AI's confidence level, which is calculated by the probability thresholds, and (c) a saliency map (calculated by the Grad-CAM++ algorithm [52]) that highlights the spatial support for the mitosis class in the reference image on the left.

- **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 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 4.7c) or the highest Ki-67 proliferation index (Figure 4.7d). 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 4.7a) or the highest Ki-67 proliferation index (Figure 4.7b). Using this tool, pathologists can locate foci of highly-mitotic areas that the highest region sampling might miss.

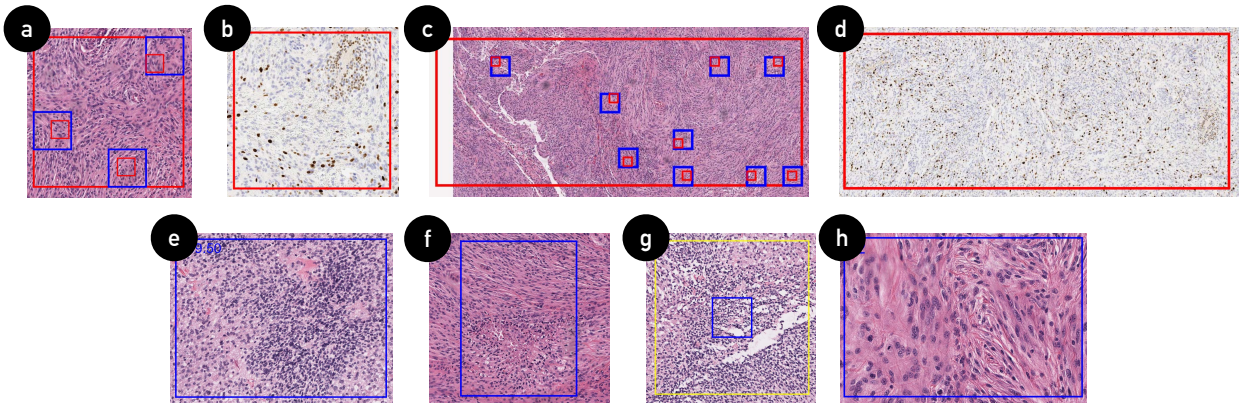


Figure 4.7: Selected pieces of sampled evidence: (a) a highest focal region sampling result of mitotic count on H&E slide (red box, 1HPF), the small blue frames indicate the rough positions of detected mitoses, and the smaller red boxes in the blue frames mark the positions of mitoses (that are shown 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).

Pathologists can go beyond the sampled areas and navigate the high-heat areas using heatmaps generated for the whole slide. 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 [43], 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" [102]) 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.

4.5 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.

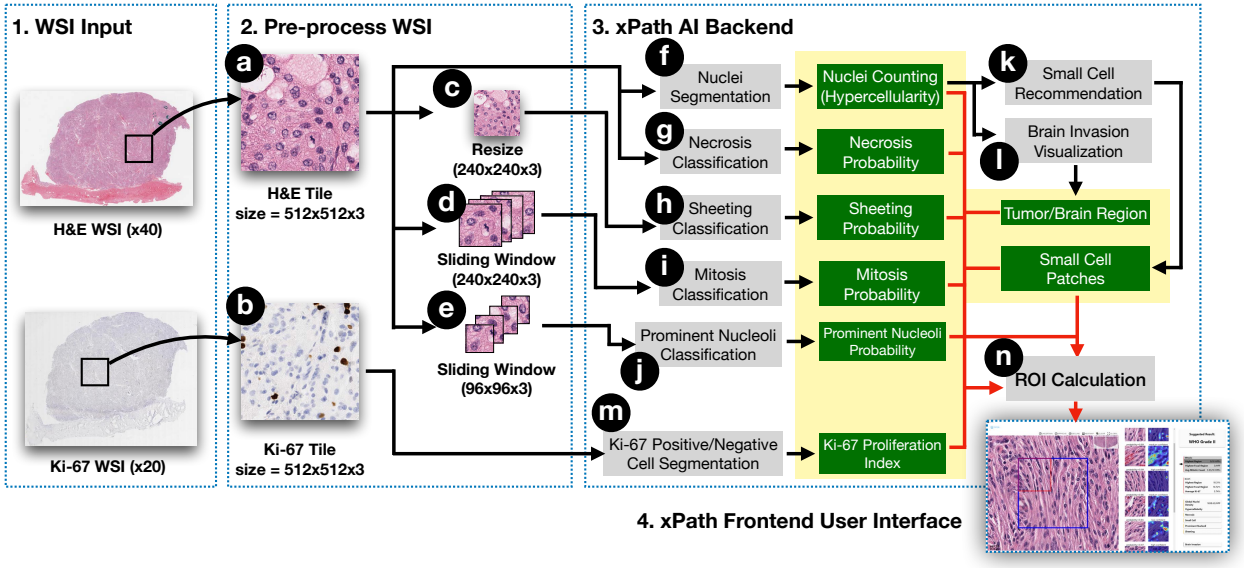


Figure 4.8: Data processing pipeline of xPath: *(i)* xPath takes H&E and Ki-67 whole slide images (WSIs) as input. *(ii)* For each WSI, xPath uses a sliding window method to acquire (a) H&E and (b) Ki-67 tiles; Furthermore, each H&E tile is processed with (c) resizing, (d) sliding window ($240 \times 240 \times 3$), 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.

4.5.1 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 4.8f), necrosis probability (Figure 4.8g), sheeting probability (Figure 4.8h), mitosis (Figure 4.8i), prominent nucleoli (Figure 4.8j), and Ki-67 proliferation index (Figure 4.8m)). 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 4.8k). Furthermore, the brain invasion was visualized by classifying the brain *vs.* tumor regions according to the nuclei count (Figure 4.8l). 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.

4.5.2 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 = $\sim$ 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 has a size of $\sim$ 16.1 GB. It consists of four training and testing sets, covering the four pathology patterns (as shown in Table

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

Table 4.2: 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.

4.2).

To train the models, for each dataset, we further randomly selected a subset of the training set to be the validation set, and utilized it to determine the optimal thresholds. The validation sampling rates for mitosis, necrosis, sheeting, and prominent nucleoli datasets are 25%, 30%, 30%, and 30%, respectively. Specific thresholds were decided by the maximum F1 scores achieved by each model in the validation set.

- **Mitotic Count (Classification)**. xPath uses an EfficientNet-b7 model [193] to identify mitosis (Figure 4.8i). A 240×240 tile with a prediction probability > 0.78 is counted as positive. xPath further applies a non-maximum suppression technique to post-process the overlapping positive tiles. The mitotic distribution of the slide is calculated by merging the results from each $512 \times 512 \times 3$ H&E patch.

- **Ki-67 Proliferation Index (Semantic Segmentation)**. xPath uses a pre-trained Cycle-GAN model [82] to detect both Ki-67 positive and negative nucleus (Figure 4.8m). Given a $512 \times 512 \times 3$ Ki-67 patch as the observation region, the Ki-67 proliferation index is calculated as $\frac{\text{positive-count}}{\text{positive-count+negative-count}} \times 100\%$.

- **Hypercelluarity (Semantic Segmentation)**. xPath uses a pre-trained deep neural

network (*i.e.*, HoVer-Net [86])to segment and count the number of nuclei in a $512 \times 512 \times 3$ H&E patch (Figure 4.8f).

- **Necrosis (Classification)**. xPath uses an EfficientNet-b5 model to judge whether a $512 \times 512 \times 3$ H&E patch contains the necrosis pattern (Figure 4.8g). A patch is considered necrosis-positive if prediction probability > 0.74.

- **Small Cell (Rule-Based Recommendation)**. xPath applies the rules for recognizing small cell patterns: selecting and recommending the top-10 $512 \times 512 \times 3$ H&E patches with the highest nuclei count within each slide (Figure 4.8k). For each recommended patch, if it has >125 nucleus/patch, then xPath includes it in the recommendation.

- **Prominent Nucleoli (Classification)**. Similar to mitosis classification, xPath uses an EfficientNet-b0 model to classify prominent nucleoli (Figure 4.8j). To avoid false-positive cases influencing the result, only tiles that have >0.9 prediction probabilities are counted as positive[5]. xPath counts positive tiles in each $512 \times 512 \times 3$ patch for calculating the distribution of prominent nucleoli.

- **Sheeting (Classification)**. xPath uses an EfficientNet-b1 model [193] to classify whether the patch includes a sheeting pattern (Figure 4.8h). A sheeting patch is called as positive if its prediction probability is >0.52.

- **Brain Invasion (Classification)**. xPath outlines the brain invasion pattern by classifying whether a given $512 \times 512 \times 3$ H&E patch is tumor, brain, or background (Figure 4.8l). If the tumor cells are invading the normal brain tissues, it can be seen with a heatmap visualization of tumors *vs.*brain areas. Because meningioma is a high-cellular tumor, xPath classifies tumor patches with the following rule: *(i)* patches

[5]We choose precision rather than recall in the prominent nucleoli classification because unlike mitosis, this criterion is justified by the presence of cell *clusters* that have prominent nucleolus, and missing one or a few detections would not significantly influence the overall result.

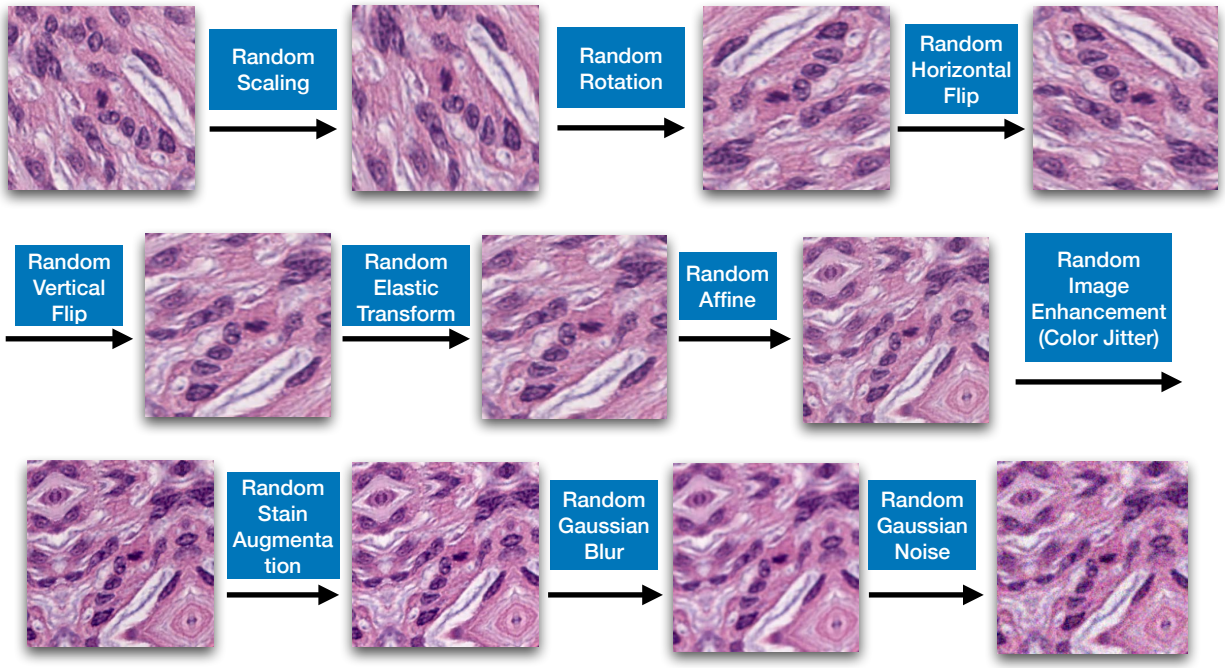


Figure 4.9: Illustration of the data augmentation pipeline used for model training to classify mitosis, necrosis, prominent nucleoli, and sheeting.

that have >55 nucleus in each H&E patch are counted as the tumor; *(ii)* patches that have [10,55] nucleus are counted as the brain; *(iii)* otherwise, count as the background.

Four deep learning models for mitosis, necrosis, sheeting, and prominent nucleoli, share the same online data augmentation strategy, which includes scaling, rotation, horizontal flip, vertical flip, elastic transform, affine, color jitter, stain augmentation [194], Gaussian blur, and Gaussian noise (see Figure 4.9).

We used the same set of hyperparameters to train the four models: initial learning rate=0.05 with decaying factor=0.3 every 15 epochs, SGD optimizer with momentum=0.9, weight decay=10^{-4}, 130 epoch with early stopping, cross-entropy loss.

The training process was performed with the corresponding datasets on a CentOS 7 server, with Intel Xeon W-2133 CPU, 104 GB memory, and two Nvidia RTX-2070 graphics cards.

4.5.3 Rules for Generating ROIs

After the AI models had processed each tile, xPath calculated the ROIs using a following set of rules (Figure 4.8n):

- A hypercellularity ROI is defined as a cluster that has >150 nucleus per tile;
- A necrosis ROI is defined as a cluster that has probability >0.7 from necrosis classifier;
- Small cell ROIs are defined jointly by *(i)* top 10 tiles that have the most nuclei count over the slide, *and (ii)* >125 nucleus/tile;
- A prominent nucleoli ROI is defined as a cluster that has >4 prominent nucleolus per tile;

Specific numbers mentioned above were decided jointly with our empirical experience and discussion with our pathologist collaborators. The areas of hypercellularity, necrosis, and prominent nucleoli ROIs are calculated by the DBSCAN clustering algorithm [70]. If no matching ROI is found on one slide, xPath's corresponding evidence list would be displayed as "no evidence found".

4.6 Constructing Mid-Level Evidence for the Mitosis Criterion

xPath includes three components in the mid-level samples for mitosis to add the explainability, including AI's probability, AI's confidence level, and a saliency map that shows spatial support for the mitosis class. We describe the implementation as follows:

- **AI's probability**: xPath applies a softmax function to the model output for calculating the AI's probability;
- **AI's confidence level**: xPath assigns each piece of evidence into three confidence levels (*i.e.*, high-, mid-, and low-confidence). For implementation, xPath identifies

the confidence levels by thresholding AI's probability according to the precision scores from validation set:

  - **If** probability > 0.9, **then** classify as high-confidence (precision$\geq$0.94);
  - **Else if** probability $\in [0.85, 0.9]$, **then** classify as mid-confidence ($0.87 \leq$ precision <0.94);
  - **Else**, classify as low-confidence ($0.76 \leq$precision<0.87)

- **Saliency map**: xPath uses the Grad-CAM++ algorithm [52] to generate the saliency map. Specifically, the 30^{th} layer of the EfficientNet-b7 model is selected and extracted for the calculation.

4.6.1 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 4.10. 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.

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.

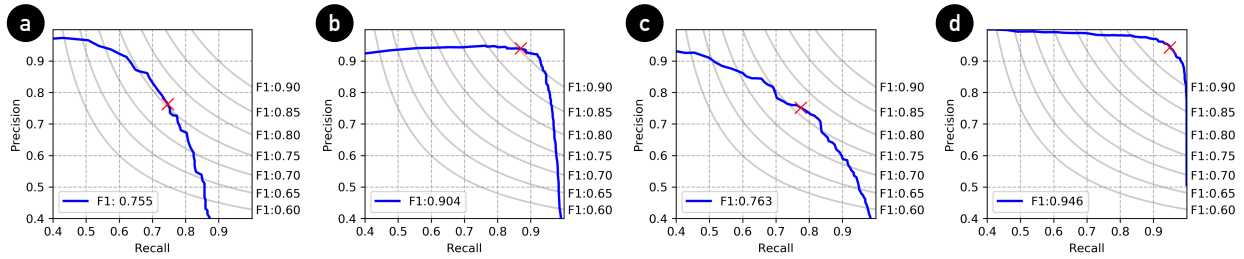


Figure 4.10: Classification performance for (a) mitosis, (b) necrosis, (c) prominent nucleoli, (d) sheeting. The solid blue lines in each sub-figure illustrate the Precision-Recall curves of each model. The red crosses indicate the performance achieved by the models using the thresholds that maximized 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.

For the brain invasion, most pathologists felt it was faster to examine it manually and did not rely on AI's recommendations.

4.7 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 [190, 119]. 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[6] and xPath. In this study, we referred to the traditional interface as system 1 and xPath as system 2 to avoid biasing of participants. The main research questions are:

[6]`https://computationalpathologygroup.github.io/ASAP/`. This tool was selected because it is open-source and has gained popularity in the digital pathology research domain [132].

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 [211, 40]; 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 [21, 39]. 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 (also shown in Chapter 2). 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.*, comprehensive-

ness, 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.**

4.7.1 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 ($\geq$ 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, $\geq$ PGY-3), and one junior resident (JR, PGY-2). The demographic information of the participants is shown in Table 4.3. 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.

ID	Occupation	Years of Experience	Frequency of Seeing WSIs	Case 1	Case 2	Case 3	Case 4	Case 5	Case 6
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 4.3: Demographic information and arrangements of the participants in the work sessions. 'Case 1' – 'Case 6' 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 4.1.)

4.7.2 Test Data

We asked our pathologist collaborators in a local medical center to select 18 meningioma slides and scan them to WSIs[7] 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 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 4.11) 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]... which include eleven H&E WSIs (scanned in x400), and seven Ki-67 WSIs (scanned in x200).

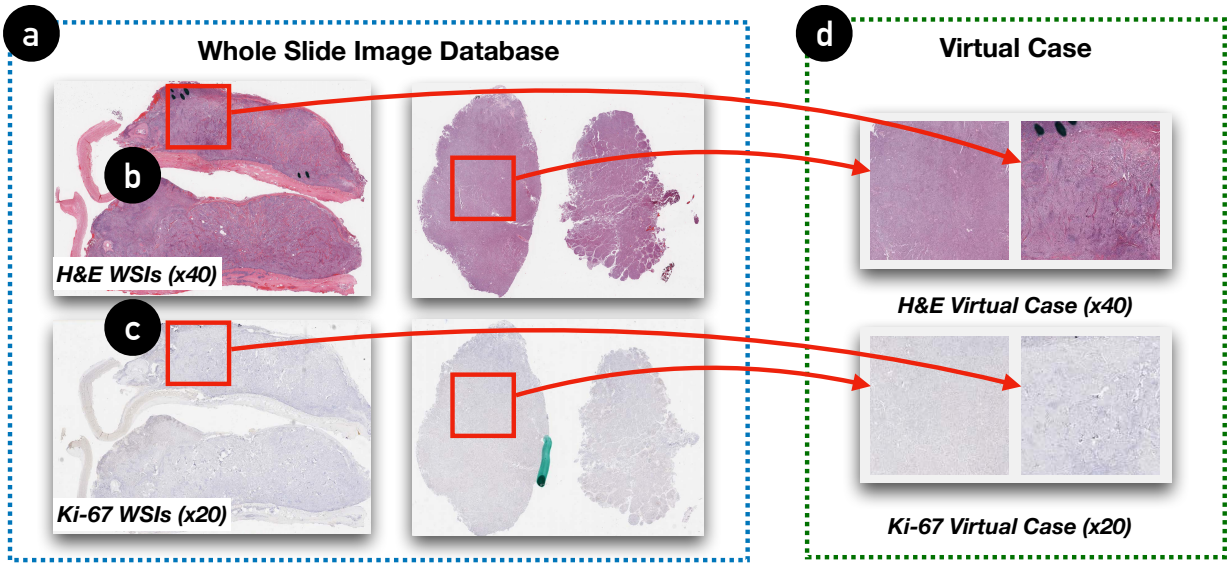


Figure 4.11: We used the 'virtual cookie cut' technique to generate the tests cases. Specifically, we first collected (a) pairs of H&E (in x400) and Ki-67 (in x200) WSIs. Then, we generated 'virtual cuts' by (b) selecting 30,000×30,000-pixel regions in H&E WSIs, and (c) 15,000×15,000-pixel regions from 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.

4.7.3 Task and 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 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[8]. 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 4.7.2); *(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.

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

4.7.4 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 4.4) with seven-point Likert questions following [43, 117, 100]. 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.

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 4.4, W1), and the effectiveness of the system in reducing the workload (Table 4.4, 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 4.4, C1). For **H3b**, we asked them to rate the explainability of xPath only since the traditional interface did not provide AI detections (Table 4.4, E1). For **H3c**, we asked participants to rate the integrability of both systems (Table 4.4, 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 4.4, T1), *(ii) How confident do you feel about the accuracy of your diagnoses using the system?* (Table 4.4, 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 4.4, F1), and also let the participants rate the overall preference of system 1 *vs.* system 2 (Table 4.4, F2).

4.8 Results and 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.

4.8.1 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 [199, 188] 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**).

4.8.2 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 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)

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 4.4: 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.

Regarding the perceived effort (**H2b**), participants reported significantly less effort (Table 4.4, W1, xPath: μ=0.91, ASAP: μ=3.67, p=0.002) and a stronger effect on reducing the workload (Table 4.4, 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)

4.8.3 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 4.4, 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 4.12e). 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)

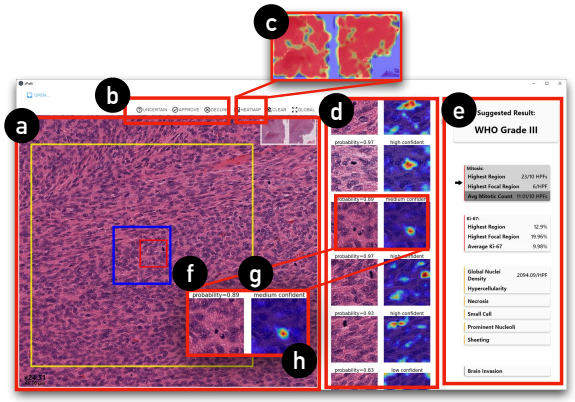
For the explainability dimension (**H3b**), xPath obtained an average rating of 5.58/7 (Table 4.4, E1). In general, participants could understand the logical relationship between the evidence and the suggested grading (global explainability). They also gave a high helpfulness rating (6.00/7, Figure 4.12d) for the list of evidence provided by xPath. However, participants gave lower ratings on the probability (3.83/7, Figure 4.12f) and the confidence level (3.92/7, Figure 4.12g) 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 4.12h). 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 4.14a) —

> *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 4.4, I1, xPath: μ=5.91, ASAP: μ=4.17, p=0.006). Specifically, participants were able



		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)

Figure 4.12: Participants' helpfulness ratings of each component in xPath. Each letter-labeled component in the right table corresponds to the marked part on the left.

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 4.4, F1, xPath: μ=6.42, ASAP: μ=3.75, p=0.002). Overall, 9/12 of the participants "to-

tally" preferred xPath over the traditional interface, while 3/12 "much more" preferred xPath (Table 4.4, 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.

4.8.4 Recurring Themes

We analyzed the video recordings of the work sessions in a similar approach as described in Section 4.3.2. 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.

4.8.4.1 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' [51] 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.

4.8.4.2 xPath's top-down workflow with hierarchical explainable evidence enables pathologists to navigate between high-level AI results and 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)

4.8.4.3 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)

4.8.4.4 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.

4.9 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.

4.9.1 Limitations and Future Improvements

We conclude the following limitations of our current work:

- 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 chapter 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 4.14) confuse some participants because they can not highlight cells appropriately;

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

4.9.1.1 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 4.13). Furthermore, the limited test cases generated for xPath's work sessions might not

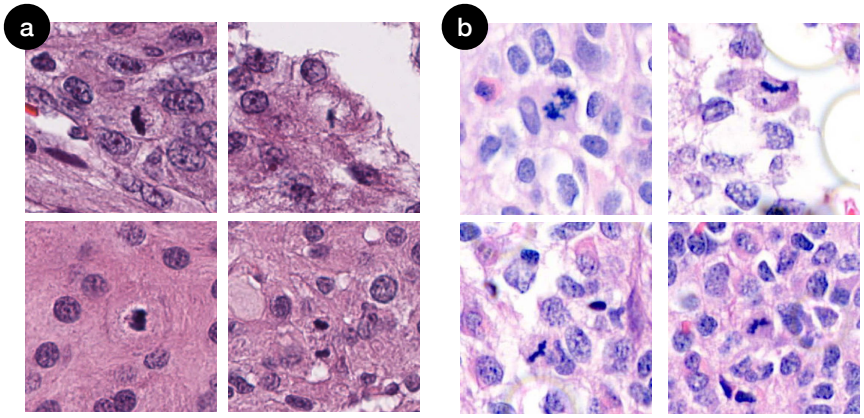


Figure 4.13: Mitoses from meningiomas (in x400), scanned by (a) the medical center in this study and (b) a different medical center. The difference in appearance is caused by the difference in processing procedures and scanners used.

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 [83]. 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 [73], 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.* [174]. 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.

4.9.1.2 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).

4.9.1.3 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 4.14, the failed saliency maps showed either scattered attention across the evidence (Figure 4.14a), or concentrated attention at the wrong place (Figure 4.14b). 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 [9].

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 [103]) to evaluate the quality of explanations.

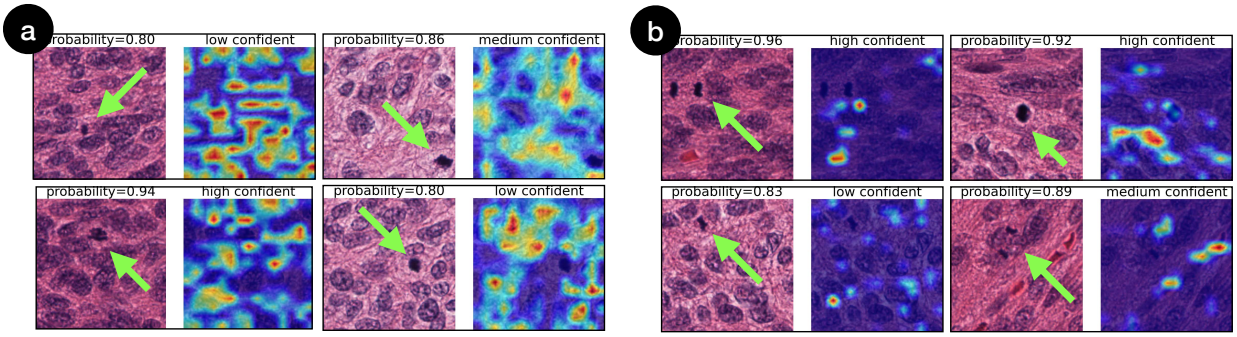


Figure 4.14: Examples of failure explanation cases, where the saliency shows (a) scattered attention across the image or (b) misleading hot spots. The green arrows point to the location of a mitosis figure marked by a human pathologist.

4.9.2 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) [137]). Here, we provide design recommendations for future physician-AI systems.

4.9.2.1 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.

4.9.2.2 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 [139] and cancer scoring [109], where the evidence is accumulated to arrive at a diagnosis.

4.9.2.3 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 [51], 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.

4.9.3 On Integrating AI into Pathologists' Workflow

4.9.3.1 How has AI improved pathologists' diagnoses in xPath?

Similar to previous human-AI collaborative research in medicine [76, 127], we discovered that using AI might improve pathologists' diagnosis quality. In pathology, the AI can "efficiently, systematically, exhaustively" analyze the entire whole slide image [188]. 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 4.3), 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 [149, 174] 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.

4.9.3.2 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 [26, 210]. 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 [39]. 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 [174].

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 [43].

4.9.3.3 How should AI systems build trust for pathologists?

Previous HCI research advises informing doctors of AI's capabilities and limitations to gain trust [44]. For example, Sendak *et al.* created a "model fact sheet" inspired by pharmaceutical drug labels to inform doctors of AI details [182]. 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 [228]), while others build inherently interpretable models (*e.g.*, concept bottleneck models [123]). 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.

4.10 Conclusion

In this chapter, 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.

CHAPTER 5

Fostering Appropriate AI Reliance in Pathology Decision-Making

This chapter is based in part on the following publication:

Hongyan Gu, Chunxu Yang, Shino Magaki, Neda Zarrin-Khameh, Nelli S. Lakis, Inma Cobos, Negar Khanlou, Xinhai R Zhang, Jasmeet Assi, Joshua T Byers, Ameer Hamza, Karam Han, Anders Meyer, Hilda Mirbaha, Carrie A Mohila, Todd M Stevens, Sara L Stone, Wenzhong Yan, Mohammad Haeri, and Xiang 'Anthony' Chen "Majority voting of doctors improves appropriateness of AI reliance in pathology." *International Journal of Human-Computer Studies* (2024): 103315.

5.1 Introduction

Although J.C.R. Licklider introduced the concept of 'man-computer symbiosis' in 1960 [130], it was not until the last decade that this vision became a more promising reality [116]. By 2023, Artificial Intelligence (AI) has been increasingly discussed to augment humans in critical tasks [31, 192, 87]. Especially in the medical domain of pathology, AI has been showcased to increase doctors' accuracy and speed [133, 131, 201, 16], consistency [19, 200], and confidence [90]. However, because pathology AI was often trained from a limited dataset its performance varied while being applied to data from new patients and hospitals [187, 11, 89]. As such, it is critical for pathologists to develop appropriate reliance while collaborating with AI, *i.e.*, to appropriately accept correct AI recommendations and reject the wrong ones.

Although there is a lack of data in pathology, research in the general domain has explored methodologies to develop appropriate reliance, focusing on reducing humans' over-reliance on AI (*i.e.*, enhancing humans' ability to reject wrong AI recommendations). Strategies, including the cognitive forcing function [39] and altering the interaction speed [158, 166, 126], have shown promising results. Additionally, effective onboarding [159] and improving AI literacy [135] were recommended and can be achieved by informing users of AI details [44, 113, 114]. However, incorporating these methods into routine medical practice presents challenges: Cognitive forcing functions could drive medical practitioners to develop algorithm aversion, leading them to reject AI recommendations even when they were correct [68, 76]. Moreover, previous studies have reported that the improvements in task accuracy with enhanced AI literacy were marginal [124, 128].

Another popular approach aims to employ explainable AI (XAI) to reduce over-reliance [42, 124, 227, 22]. However, the efficacy of XAI is countered in part by the cognitive effort for understanding these explanations [202]. Adding XAI-related content might increase doctors' cognitive burden, possibly causing them to overlook XAI. Therefore, there remains a pressing need for alternative strategies to foster appropriate AI reliance in medical applications.

By reviewing pathologists' decision-making workflows, we found that the critical decisions were usually determined through a combined judgment among multiple doctors [32]. The underlying intuition was that a group of pathologists might produce safer and more rational judgments while working together [32]. In the context of AI, recent studies have employed majority voting among pathologists' AI-assisted decisions to collect annotations for datasets [30, 12]. However, there is a lack of empirical evidence supporting that such a majority voting approach would enable appropriate reliance.

This research aims to provide the validation of the majority voting on enabling the appropriate AI reliance in pathology decision-making, with a focus on a visual search task of detecting "mitosis," a critical histology pattern for tumor grading [58, 146]. 32 medical professionals in pathology from ten institutions participated in a multi-stage user study, where

they detected mitoses manually, first, and with AI assistance after a wash-out period. Here, the majority voting decisions were synthesized according to the AI-assisted decisions from an odd number of randomly-selected pathologist participants. Two metrics were employed to measure the appropriateness of AI reliance: "relative AI reliance" and "relative self-reliance" [180]. The result showed that the majority voting decisions from as few as three pathologists showed significantly higher relative AI reliance ($\sim$ 9% increase) and relative self-reliance ($\sim$ 31% increase), compared to one pathologist collaborating with AI, respectively. The precision and recall of majority voting decisions also increased: Those from three AI-assisted pathologists could achieve a mean precision of 0.902 and a recall of 0.843. As a comparison, the mean precision and recall for one-pathologist-AI collaboration were 0.824 and 0.817, respectively. Furthermore, the majority voting decisions could also have a higher chance of achieving super-AI performance in the recall.

5.1.1 Contributions

This research showcases that majority voting can enable appropriate AI reliance for pathology decision-making. Throughout a multi-institutional study amongst 32 pathology professionals, this research presents the effectiveness of majority voting in a high-stakes medical task, which can ultimately benefit patient management. This signifies a transformation from the traditional one-human-AI collaboration to harnessing group decision-makings of AI-assisted medical professionals. While our primary focus has been on pathology, we envision that the insights of this study can have broader implications for leveraging collective human-AI decision-making in other high-stakes visual search tasks, such as detecting explosives from X-ray scans or disaster assessment from satellite imagery for emergency response efforts.

5.1.2 Sample Selection and Mitosis Ground Truth Acquisition

Meningioma specimens were collected from a local hospital after receiving ethics approval. These specimens were digitized into 19 digital slides with an Aperio CS2 Scanner (Manufacture: Leica, Germany). A specialist pathologist examined these slides and selected 51 regions of interest (ROIs) based on predefined criteria (Section 5.1.2.1). Each ROI has a dimension of $1,600 \times 1,600$ pixels ($400 \times 400 \mu$m), with one example shown in Figure 3.2(a). This image dimension matches the field-of-view under the $40\times$ objective lens in light microscopy, which can reduce the mental effort for pathologists to adapt to the digital interface.

As for collecting the mitosis ground truth, two residents independently annotated all 51 images initially. Next, a third specialist pathologist reviewed these initial annotations and provided a final decision. To ensure the accuracy of the ground truth, the three doctors referred to the results of an additional antibody test (the Phosphohistone-H3 immunohistochemistry test, a mitosis indicator usually used in medical research [67, 77], Figure 3.2(b)) in the ground truth annotation process.

Within the 51 selected ROI images, three were selected for the tutorial, leaving the rest 48 for testing purposes. The 48 test images have 88 mitoses in total. The count of mitoses per image varies between zero and six, which can cover the majority of mitosis prevalence in a single ROI in meningiomas.

5.1.2.1 Criteria for Selecting Test ROI Images for the User Study

Each ROI image has the size of $1,600 \times 1,600$ pixels (equivalent to $400 \times 400 \mu m$), which has the same size of one High-Power Field under the $40\times$ objective lens with the light microscopy. The 51 test ROI test images were selected from the test WSIs during the AI development, according to the following five criteria.

1. HPFs with no or minimal out-of-focus area.

2. HPFs with at least 50% tumor content.

3. Mitotic figures at different stages (*i.e.*, prophase, metaphase, and anaphase-telophase) should be present.

4. Very few atypical mitotic figures in the collection are allowed.

5. The number of mitotic figures ranges from none to 6 to represent mitotic counts that may be seen in tumor grades 1 – 3. HPFs with significant staining issues and tissue folding/wrinkling artifacts were removed from the collection.

5.1.3 Experience Level of Pathologists

In the United States, pathology professionals can be classified into four levels based on their training progress and experience [81]:

1. A **medical student** is currently receiving medical education.

2. A **resident** has earned their Medical Doctor or an equivalent degree and is in post-graduate residency training.

3. A **general pathologist** has completed their residency training and holds general board certification in pathology.

4. A **specialist pathologist** has received/ is undergoing further training in a sub-specialty area (in this study, neuropathology) after becoming certified as a general pathologist.

Regarding familiarity with the mitosis detection task, specialist pathologists are expected to have the highest level because of their sub-specialty training. General pathologists should have a moderate familiarity, having acquired their general board certification. As for residents and medical students, their familiarity depends on their exposure during rotations and

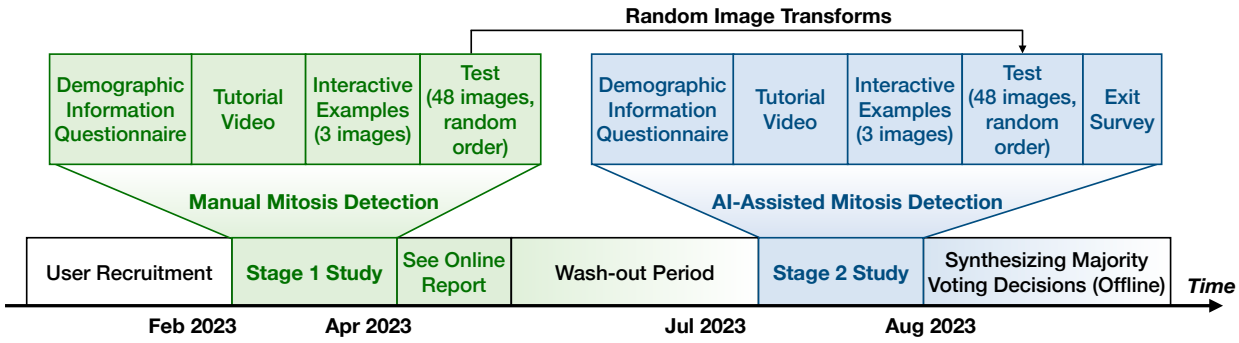


Figure 5.1: Organization of the user study.

any subsequent training they have received. In this study, 32 medical professionals from ten institutions participated, covering all four aforementioned categories.

5.2 User Study

An online user study was conducted under the Institutional Review Board approval of the University of California, Los Angeles (IRB#21-000139). The user study has two major stages (Figure 5.1): *(i)* **Stage 1** (February 2023 – April 2023): participants performed the mitosis detection task in 48 test images manually; *(ii)* **Stage 2** (July 2023 – August 2023): participants detected mitoses in the same 48 images with AI-assistance. This sequential arrangement follows previous work [180], and was designed to investigate potential shifts in pathologists' decisions influenced by AI. The majority voting decisions were synthesized offline after the stage 2 responses had been collected. The main research questions are:

- **RQ1**: How did pathologists use AI and XAI while performing the "mitosis detection" task?
- **RQ2:** How does the majority voting mechanism influence the appropriateness of AI reliance compared to one pathologist collaborating with AI?
- **RQ3**: Is the majority voting mechanism more likely to achieve complementary team

performance compared to one pathologist collaborating with AI?

5.2.1 Participants

Participants were recruited through sending emails to the mailing list and snowball recruitment. As a result, 32 participants submitted their responses in both stages of the user study, including 12 specialist pathologists (*i.e.*, neuropathologists or neuropathology fellow), 6 general pathologists, 10 pathology residents, and 4 medical students[1]. 18/32 participants were from **Institution #1**(I1), 5/32 from I2, 2/32 from I3, and the remaining 7/32 were each from a different institution. Their demographic information was summarized in Table 5.1.

Note 1 Did not activate AI for over 45/48 images at Stage 2 study. Considered as non-AI users and excluded from the all analyses.

Note 2 Years of experience (YoE) not applicable for the medical student. To ensure their familiarity with the mitosis detection task, all medical student participants underwent a 45-minute training session overseen by a specialist pathologist before participating.

The demographic information of participants is shown in Table 5.1.

5.2.2 Study Procedure

Stages 1 and 2 of the user study were conducted in an unmoderated manner. At each stage, each participant joined online with their computers at the recommended display settings. The study of each stage consisted of the following parts (Figure 5.1):

1. **Demographic information**: Participants filled in a demographic information questionnaire.

[1]The four medical student participants underwent a 45-minute training session overseen by a specialist pathologist before participating, to ensure their familiarity with the mitosis detection task.

Table 5.1: Demographic Information of the Participants

Index	Experience Level	Institution	YoE	Note
1	Specialist Pathologist	I1	5–10	
2	Specialist Pathologist	I2	5–10	
3	Specialist Pathologist	I3	>10	
4	Specialist Pathologist	I4	5–10	
5	Specialist Pathologist	I1	5–10	
6	Specialist Pathologist	I2	>10	See **Note 1**
7	Specialist Pathologist	I5	>10	
8	Specialist Pathologist	I6	>10	
9	Specialist Pathologist	I2	5–10	
10	Pathology Resident	I1	2–5	
11	Specialist Pathologist	I7	5–10	
12	Pathology Resident	I2	2–5	
13	Pathology Resident	I1	2–5	
14	Pathology Resident	I1	2–5	See **Note 1**
15	Specialist Pathologist	I8	5–10	
16	General Pathologist	I2	5–10	
17	General Pathologist	I9	5–10	
18	General Pathologist	I1	5–10	
19	General Pathologist	I1	5–10	
20	General Pathologist	I1	>10	See **Note 1**
21	Pathology Resident	I1	2–5	
22	Pathology Resident	I3	2–5	
23	General Pathologist	I1	5–10	
24	Medical Student	I1	N/A	See **Note 2**
25	Medical Student	I1	N/A	See **Note 2**
26	Pathology Resident	I1	2–5	
27	Specialist Pathologist	I10	>10	
28	Pathology Resident	I1	2–5	
29	Pathology Resident	I1	2–5	
30	Pathology Resident	I1	2–5	
31	Medical Student	I1	N/A	See **Note 2**
32	Medical Student	I1	N/A	See **Note 2**

2. **Tutorial**: Participants saw a tutorial video describing how to participate, followed by an interactive tutorial of three example images. No AI details were revealed to participants.

3. **Test**: Participants examined the 48 images without (stage 1) or with (stage 2) AI assistance. Their task was to detect and report mitoses from these images with their threshold of daily practice.

Two methods were introduced to reduce the learning effect of participants in the stage 2:

- **Random image transforms**: Including random flipping (vertical and/or horizontal) and random rotation (randomly chosen from {0°, 90°, 180°, and 270°}). For instance, the image shown in Figure 5.2(b) was rotated 270° anti-clockwise from that in Figure 5.2(a).

- **Wash-out period and ground truth blinding**: After completing stage 1, participants received personalized online report documents highlighting disagreements between their mitosis reportings and the ground truth. After two weeks, they were prevented from accessing these online documents. Next, after a wash-out period of three months, they were invited to participate in the stage 2 study.

5.2.3 User Interfaces and Key Features

For each stage, we deployed an interface online to enable participants to examine the images and report mitoses.

5.2.3.1 Stage 1: Manual Mitosis Detection

This interface only showed participants the images and logged their interactions (Figure 5.2(a)). If the user found a mitosis, they could left-click on where it resided to leave a

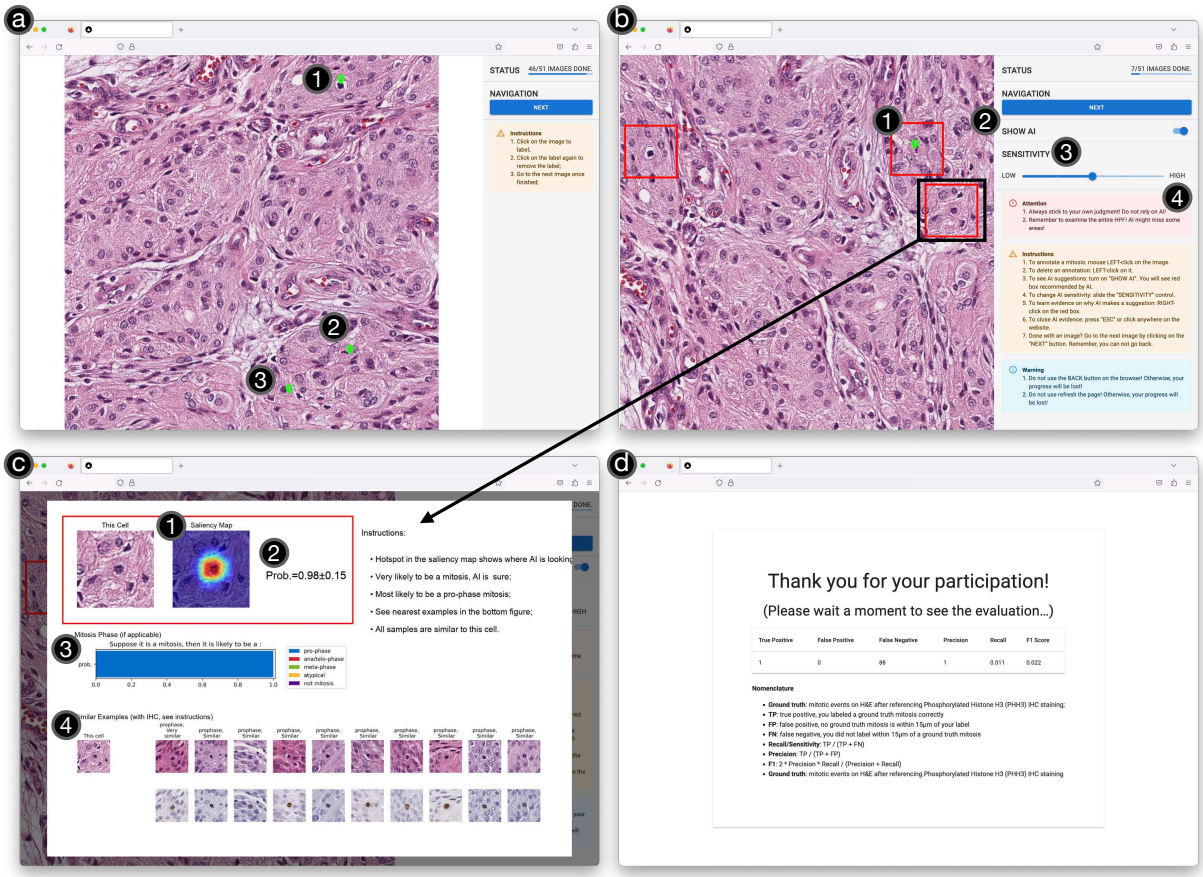


Figure 5.2: Screenshots of the mitosis study websites: (a) The manual mitosis detection website in the stage 1 study. The user could left-click on the image to leave a mark for each mitosis detected (① – ③). (b) The AI-assisted mitosis detection website in the stage 2 study. The interface added ① the AI recommendation box; ② "Show AI" switch, where the user could toggle on/off AI recommendations; ③ "AI Sensitivity" slider, where the user could adjust the sensitivity of AI based on their preference; ④ a warning message to remind users not relying on AI. (c) The website in stage 2 also provided an XAI evidence card for each AI recommendation. Each XAI evidence card included ① a saliency map; ② confidence level, including a probability score and a trust score; ③ a bar plot for subclass probability; and ④ similar examples. (d) After the user finishes examining all images, an evaluation page will inform the performance metrics to the participant.

mark (Figure 5.2(a) ① – ③). The user could go to the next image after examining one. However, they could not return to the previous image to ensure a precise measurement for time consumption. After all images were examined, a status page (Figure 5.2(d)) was displayed to inform the participant of the performance of their mitosis detection.

5.2.3.2 Stage 2: AI-Assisted Mitosis Detection

The AI model used in this stage was an EfficientNet-b3 Convolutional Neural Network (CNN), trained from a meningioma mitosis dataset [193, 88]. The website displayed AI mitosis detections through recommendation boxes (Figure 5.2(b)). Additionally, following previous works, we included four components to mitigate the negative influence of improper AI reliance:

- **Warning messages**: A "black-box" style[2] warning message was presented in the tutorial video, suggesting that the users should always rely on their judgments. The message was also shown in a highlighted box on the website (Figure 5.2(b) ④).
- **XAI**: Each AI recommendation was accompanied by an evidence card which attempts to provide XAI assistance [161]. The user could right-click on the AI recommendation box to see the XAI evidence card *on-demand*. Four popular XAI techniques were included following previous work [72], including:
  - **Saliency map**: Generated by GradCAM++ [52].
  - **Confidence level**: Including a probability score and a trust score [227]. The trust score was the geometric mean of noise [15] and random AI variances [79] of the AI prediction.
  - **Subclass**: A bar plot showcasing potential subclasses of the mitosis (*i.e.*, prophase, meta-phase, ana/telo-phase, atypical, and not mitosis) in this AI recom-

[2]... the highest safety-related warning assigned by the U.S. Food and Drug Administration [65].

mendation.

– **Similar examples**: A set of ten similar instances was retrieved from an annotated dataset that includes paired Hematoxylin and Eosin – immunohistochemistry staining [43].

Counterfactual explanations were not used because of the low quality of the retrieval results achieved by the our AI model.

- **Personalized AI adjustments**: The user could toggle on/off AI recommendations by interacting with the "Show AI" switch (Figure 5.2(b) ②) (*i.e.*, AI on-request, suggested by [80]) and adjust the AI sensitivity (Figure 5.2(b) ③) according to their preferences. The website provided five AI sensitivity settings for users: "lowest," 'low," "medium," "high," and "highest." A higher sensitivity would include more AI recommendations with lower probabilities.

- **Random image order**: The 48 images were presented to participants in a random order to prevent users from anchoring on AI based on their initial impressions (*i.e.*, the ordering effect [154]).

We chose not to reveal AI information to participants because of the time-consuming nature of the education process.

5.2.4 AI Training Detail and WSI-Level Evaluation Result

An EfficientNet-b3 Convolutional Neural Network (CNN) [193] was trained and evaluated based on the 19 H&E Whole Slide Images (WSIs) and their mitosis annotations, with 10/19 slides for training/validation, and 9/19 for testing. The upper 75% areas of the training WSIs were used for training, leaving the lower 25% for validation and hyperparameter tuning.

The model training process involves a multi-round activate learning strategy: for each round, both training and evaluation sets consists of patches (size=240×240) extracted from

the corresponding areas in the WSIs. In each round, the model was trained with Stochastic Gradient Descent optimizer (momentum=0.9), Cosine Annealing learning rate scheduler with warm restart (max learning rate=6×10^{-4}), with combined online uncertainty sample mining (k=1) [219] and Consistent Rank Logits loss [50], batch size 128, data augmentation as specified in [88], and 80 epochs. After each round of training, the model with best validation F1 score was applied to the train WSIs. Correspondingly, false-positive and false-negative patches were added to the training/validation set for the next round of training. The training process repeated four times when the validation F1 score stopped increasing, resulting 102,575 patches in the training set and 24,792 in the validation. The model with the best validation F1 score in the last round of training, along with the corresponding threshold, was chosen for evaluation.

For evaluation, the trained CNN was applied on the test WSIs with a window size of 240 $\times$ 240 pixels and a step size of 45 pixels. A window with a probability greater than 0.70 was regarded positive, and overlapping positive windows were removed by non-max suppression with a threshold of 0.10. For each remaining positive window, a class-activation map was calculated using GradCAM++ [52], and the centroid of each hotspot in the class-activation map was extracted as a candidate mitosis location. The CNN was applied again on these candidate locations for step-2 verification, and candidates with a step-2 verification probability greater than 0.775 (selected based on the best detection F1 score) were considered as positive mitosis detections.

Figure 5.3 shows the precision-recall curve of the CNN on the nine test WSIs. With the threshold cut-off of 0.775, the model achieved a 1,091 TP, 437 FP, and 491 FN mitoses, resulting a precision of 0.714, recall(sensitivity) of 0.690, and an F1 score of 0.702.

Table 5.2 shows the precision and recall values of the CNN on the 48 test images under the five "AI Sensitivity Settings" in stage 2 user study. The "best validation" stands for the best validation threshold used in the model development phase (0.775). And the "best validation" condition was selected for the "complementary team performance" analysis in

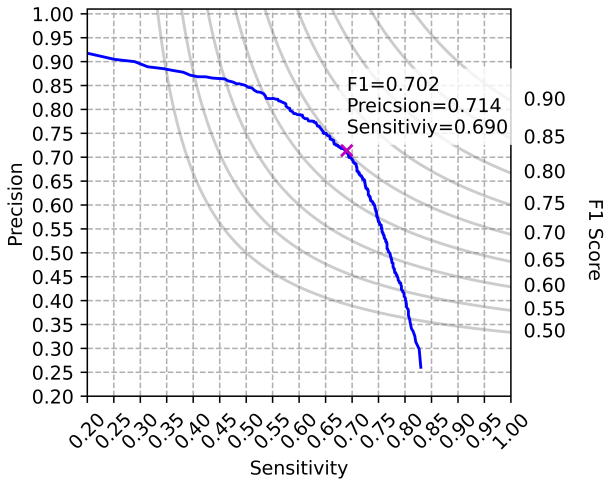


Figure 5.3: Precision-recall curve for the AI model on the nine test WSIs.

Table 5.2: Operating point selection for the "AI Sensitivity Setting" feature in the stage 2 user study.

Threshold	**Sensitivity Setting**	**Precision**	**Recall**
0.00	`Highest`	0.708	0.909
0.10	`High`	0.786	0.875
0.30	`Medium`	0.855	0.875
0.50	`Low`	0.894	0.836
0.70	`Lowest`	0.961	0.841
0.775	Best Validation	0.961	0.841

the main text of the paper.

5.2.5 Development of eXplainable AI Evidence Card

Figure 5.4 demonstrates an example of the eXplainable AI (XAI) evidence card used in the stage 2 study. For each AI detection, an XAI evidence card with four types of explanations was generated automatically. This section introduces the development and implementation of these XAI methods.

5.2.5.1 Saliency Map

The saliency map was generated by the GradCAM++ [52], as shown in Figure 5.4(a).

5.2.5.2 Confidence level

For each AI-detected mitotic figure (I), we applied the mitosis detection model 100 times, including 50 times with Dropout layers enabled (Θ_{Dropout}, stands for the random errors incurred by the model), and 50 times with added Gaussian noise as input ($(\hat{I})$, stands for the errors incurred by the noise). The probability score was calculated according to Equation 5.1. The trust score indicates the confidence of the probability score, and was calculated as the geometric mean of the standard deviations of the 50 "Dropout" predictions and the 50 "noise" predictions, according to Equation 5.2.

$$\text{Probability Score}(I) = \sqrt{\frac{\sum_1^{50} \Theta_{\text{Dropout}}(I)}{50} \times \frac{\sum_1^{50} \Theta(\hat{I})}{50}} \tag{5.1}$$

$$\text{Trust Score}(I) = \sqrt{\sigma(\Theta_{\text{Dropout}}(I)|_1^{50}) \times \sigma(\Theta(\hat{I})|_1^{50})} \tag{5.2}$$

The evidence card showed two information from confidence level

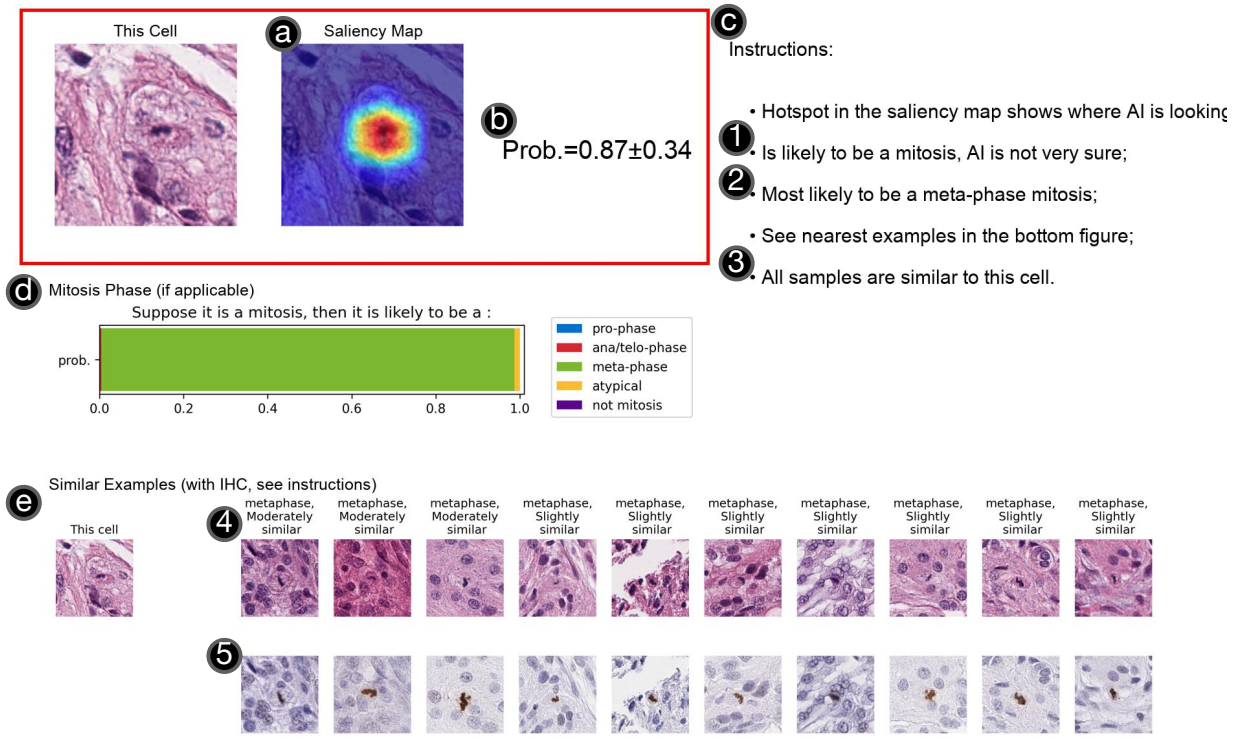


Figure 5.4: An example of the eXplainable AI evidence card, including the following components: (a) a saliency map; (b) the confidence level, including the probability score and trust score; (c) verbal descriptions of the evidence card, explaining the implications of ① the probability score and trust score, ② mitosis subclass of this detection, and ③ summary of the similarity qualities of the retrieval results; (d) stacked probability plot for the mitosis subclass; (e) similar examples, with retrieved pairs of ④ annotated H&E mitoses and their ⑤ PHH3-IHC counterparts in our database.

- Figure 5.4(b): "Prob.= (Probability Score) ± (Trust Score)".
- Figure 5.4①: Verbal descriptions of the confidence level:
  "`Is [Placeholder # 1] likely to be a mitosis, AI is [Placeholder # 2] sure.`"

Rules for the `Placeholder #1` in the verbal description:

- `very`: if probability score > 0.95;
- `[empty]`: else if probability score > 0.75;
- `moderately`: else if probability score > 0.38;
- `slightly`: else if probability score > 0.21;
- `not very`: else.

Rules for the `Placeholder #2` in the verbal description:

- `not very`: if trust score > 0.30;
- `slightly`: else if trust score > 0.25;
- `moderately`: else if trust score > 0.17;
- `[empty]`: else if trust score > 0.06;
- `very`: else.

5.2.5.3 Subclass (Mitosis Phase)

Based on the morphology, mitoses can have approximately classified into five subclasses: "prophase", "metaphase", "anaphase", "telophase", and atypical. Here, we consider "anaphase"

Table 5.3: AP and mAP of the CBIR model

Subclass	AP@1	AP@5	AP@10	AP@50	AP@100	mAP@100
Prophase	0.77	0.71	0.72	0.71	0.71	0.73
Metaphase	0.73	0.74	0.74	0.75	0.75	0.75
Ana-Telophase	0.49	0.50	0.51	0.51	0.51	0.52
Atypical	0.29	0.32	0.30	0.29	0.25	0.30
Not Mitosis	0.47	0.46	0.47	0.45	0.44	0.48

and "telophase" as one subclass – "ana-telophase" because of the difficulty in distinguishing them. All mitoses from the 19 H&E WSIs in Section 5.2.4 were labeled with corresponding subclasses. All mitoses from the 10/19 training WSIs in Section 5.2.4 were involved in fine-tuning process, with 80%/20% train/val split. We fine-tuned the EfficientNet-b3 model trained in Section 5.2.5.3 by freezing the weights of Blocks #0 – # 25. For the rest learnable parameters, the max learning rate (LR) was set at 1×10^{-4}, Cosine Annealing LR scheduler with warm restart (T_0=5 epochs, T_{mult}=2), SGD optimizer, loss from Section 5.2.4, 300 epochs. The best model with the smallest validation loss was selected as the final subclass model.

Mitoses from the rest 9/19 test slides are used for test images. The precision-recall curve for each subclass is shown in Figure 5.5.

The evidence card showed two information from the subclass AI model.

- Figure 5.4(d): A stacked-bar plot for the probability distribution of each subclass.
- Figure 5.4②: The name of the subclass with the highest probability.

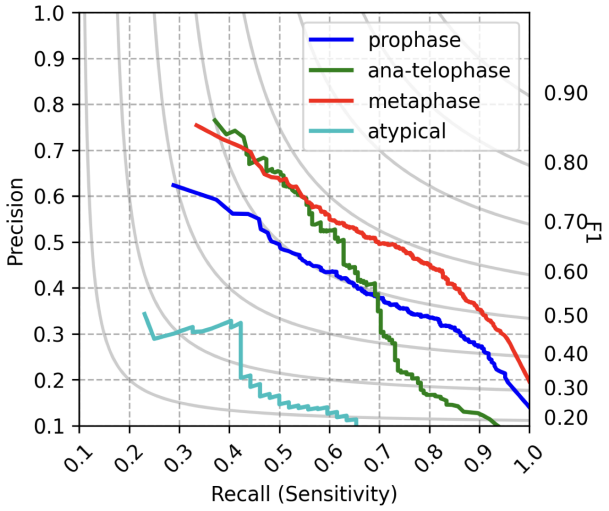


Figure 5.5: Precision-Recall curve of each subclass of "prophase", "metaphase", "ana-telophase", "atypical" mitoses.

5.2.5.4 Similar Examples

The "similar example" evidence was developed by training a Content-Based Image Retrieval (CBIR) model. The CBIR used the weight of the EfficientNet-b3 model trained in Section 5.2.5.3. Similar examples were retrieved according to the smallest $l2$ distance between the embeddings of queried sample and those of the annotated subclass database (Section 5.2.5.3), train/val fold, H&E. As for the testing, for each subclass of "prophase", "metaphase", "ana-telophase", "atypical", and "not mitosis", we randomly selected 100 samples from the test fold in Section 5.2.5.3 and calculated the average precision (AP) at 1, 5, 10, 50, and 100. The mean average precision (mAP) at 100 was also reported. The evaluation results were summarized in Table 5.3.

The evidence card showed two information from confidence level

- Figure 5.4(e): Ten most similar sets of H&E-PHH3 IHC similar examples from the annotated subclass database (train/val fold).
- Figure 5.4④: Verbal descriptions of each similar example:

"`subclass, [Placeholder # 3] similar.`"

- Figure 5.4③: Brief summary of the similarity qualities of the retrieval results: "`[Placeholder #4] samples are similar to this cell.`"

Rules for the `Placeholder #3`:

- `very`: if $l2$ distance < 50;
- `[empty]`: else if $l2$ distance < 63;
- `moderately`: else if $l2$ distance < 71;
- `slightly`: else if $l2$ distance < 110;
- `not very`: else.

`Placeholder #4` shows the number of retrieved samples that are rated with ``` ``slightly'' ``` at least. For instance, in Figure 5.4(e), all retrieved samples were rated with ``` ``slightly'' ``` at least. Therefore, the description said "`All samples are similar to this cell.`"

5.2.6 Synthesizing Majority Voting Decisions from Groups of AI-Assisted Participants

Participants' majority voting decisions were synthesized offline after collecting their responses from the stage 2 study. It consisted of two steps:

Step 1 **Random Sampling**: Mitosis reportings from an odd number k participants from stage 2 were aggregated as a group (Figure 5.6(a)). Members in a group were sampled randomly from the participant pool without replacement.

Step 2 **Majority Voting**: Mitoses candidates reported by more than half of members ($k/2$) in the group remained as the final majority voting decision (Figure 5.6(b)).

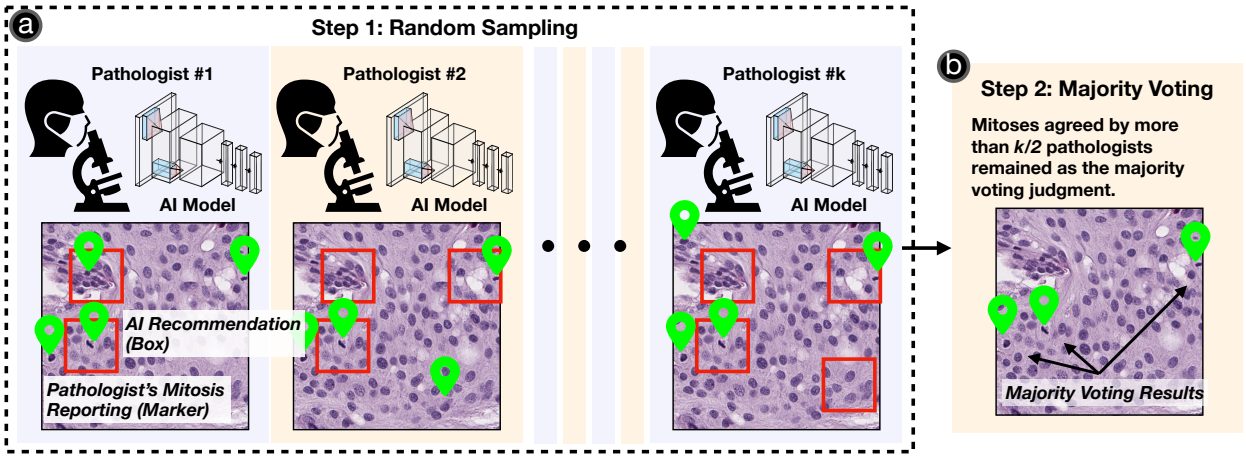


Figure 5.6: Steps for synthesizing the majority voting decisions from k AI-assisted pathologists: (a) random sampling: mitosis reportings from an odd number of k randomly-sampled, AI-assisted pathologists were collected, (b) majority voting: mitoses candidates reported by $> k/2$ pathologists remained as the final decision.

Group sizes of odd numbers $k = 3, 5, 7, \ldots, 27$ were explored. For each group size, the random sampling–majority voting processes were run 100 times for further analysis.

5.2.7 Measures and Statistics

5.2.7.1 Utilization of AI and XAI (RQ1)

We employed two metrics to measure how participants used AI assistance in the stage 2 study:

- **AI activation rate**: Indicating the percentage of the 48 test images where the AI was activated at least once (Equation 5.3).
- **AI active time percentage**: Since the participant might deactivate the "Show AI" feature, this metric represents the percentage of time when the "Show AI" feature stayed active during the entire stage 2 study (Equation 5.4).

$$\text{AI activation rate} = \frac{\sum_{i=1}^{48} \mathbb{1}[\text{“Show AI” in image}_i == \text{“On”}]}{48} \times 100\% \tag{5.3}$$

$$\text{AI activation time percentage} = \frac{\sum_{i=1}^{48} T[\text{“Show AI” in image}_i == \text{“On”}]}{\sum_{i=1}^{48} \text{Time consumption on image}_i} \times 100\% \tag{5.4}$$

Participants' utilization of XAI was measured by the following two metrics:

- **XAI activation rate** was calculated according to Equation 5.5. The number of "AI recommendations in image$_i$" was counted based on the highest sensitivity set by a participant while they examined the image$_i$. If the "Show AI" was not toggled on in an image, then it was not counted.
- **XAI activation time** was measured by the time elapsed between a participant opening and closing an XAI evidence card.

$$\text{XAI activation rate} = \frac{\sum_{i=1}^{48} |\text{XAI opened in image}_i|}{\sum_{i=1}^{48} |\text{AI rec. in image}_i| \times \mathbb{1}[\text{“Show AI” == “On”}]} \times 100\% \tag{5.5}$$

5.2.7.2 Reliance on AI (RQ2)

We used the categorization proposed by [180] to define the incidents related to the reliance. Four types of events were defined under the categorization: *(i)* correct self-reliance, *(ii)* incorrect AI reliance (over-reliance), *(iii)* correct AI reliance, and *(iv)* incorrect self-reliance (under-reliance). The criteria for judging these events were based on the true-positive (TP), true-negative (TN), false-positive (FP), and false-negative (FN) detecitons[3]. We adopted the framework in [180] for the mitosis detection task, which is summarized in Figure 5.7.

[3]A TP was defined as "there was a ground truth within 60 pixels ($15\mu m$) of a participant-reported mitosis," a TN was "no participant-reported mitoses were found surrounding a non-mitotic figure," an FP was "no ground truth was found within a 60-pixel radius of a participant-reported mitosis," and an FN was "no participant-reported mitoses were found within 60-pixel radius of a ground truth."

Correctness				Scenarios		Indication
Mitosis Ground Truth	Human Decision (Stage 1)	AI Recommendation (Stage 2)	Human-AI Decision (Stage 2)	Did human call it a mitosis?		
				Stage 1	Stage 2	
Not Mitosis	**TN**	*FP*	**TN**	No	No	Correct Self-Reliance (CSR)
Mitosis	**TP**	*FN*	**TP**	Yes	Yes	
Not Mitosis	**TN**	*FP*	*FP*	No	Yes	Incorrect AI Reliance (Over-reliance)
Mitosis	**TP**	*FN*	*FN*	Yes	No	
Mitosis	*FN*	**TP**	**TP**	No	Yes	Correct AI Reliance (CAIR)
Not Mitosis	*FP*	**TN**	**TN**	Yes	No	
Mitosis	*FN*	**TP**	*FN*	No	No	Incorrect Self-Reliance (Under-reliance)
Not Mitosis	*FP*	**TN**	*FP*	Yes	Yes	

Figure 5.7: Combinatorics for reliance incidents in the condition of one pathologist collaborating with AI (*i.e.*, one-human-AI) for the mitosis detection task. This chart is adopted from the framework described in [180].

Schemmer *et al.*further introduced two normalized metrics, Relative AI Reliance (RAIR), and Relative Self-Reliance (RSR), to represent the Appropriateness of Reliance (AoR). The RAIR relates to the under-reliance events (Eq. 5.6). And the RSR relates to the over-reliance events (Eq. 5.7). The Appropriateness of Reliance is encapsulated by the tuple of PAIR and RSR (Eq. 5.8), which can be graphically represented on a 2D chart with the RAIR on the x-axis and the RSR on the y-axis.

$$\text{Relative AI reliance (RAIR)} = \frac{\text{Correct AI Reliance}}{\text{Correct AI Reliance} + \text{Under-reliance}} \tag{5.6}$$

$$\text{Relative Self reliance (RSR)} = \frac{\text{Correct Self Reliance}}{\text{Correct Self Reliance} + \text{Over-reliance}} \tag{5.7}$$

$$\text{Appropriateness of Reliance (AoR)} = (RSR; RAIR) \tag{5.8}$$

To measure AI reliance on majority voting decisions, we also implemented the majority voting process for stage 1. To ensure a "with-in-subject" nature of the analysis, for each majority voting run for stage 2, a vis-à-vis majority voting from the same group of participants

Table 5.4: Modified definitions to measure AI reliance for the majority voting decisions synthesized from a group of k pathologists.

Items	Majority Voting Decisions (Group Size=k)
Human Decision (stage 1)	Majority voting results based on the stage 1 decisions from k participants
AI Recommendation (stage 2)	For each image, AI recommendations under the highest sensitivity set by more than $k/2$ of participants while they were seeing the ROI
Human-AI Decision (stage 2)	Majority voting results based on the stage 2 decisions from the same k participants

in stage 1 was conducted. The definitions of "human decisions," "AI recommendations," and "human-AI decisions" were adjusted to fit the majority voting condition and are summarized in Table 5.4.

Because participants might employ different AI sensitivity settings in stage 2, the random sampling process to formulate groups was also adopted with regard to each participant's AI sensitivity setting: for the AI reliance analysis, the k pathologists were exclusively drawn from the subset of pathologists who majorly set the same AI sensitivity, which ensured the AI conditions among all group members were similar.

To study **RQ2**, we compared five conditions: one pathologist collaborating with AI (*i.e.*, one-human-AI collaboration), and majority voting for the four group sizes (k=3,5,7,9). For each criterion of RAIR and RSR, a Kruskal–Wallis test was first applied to show significance among these five conditions. A post-hoc Dunn's test with Bonferroni correction was then used to test pair-wise significance. Appropriateness of Reliance scatter plots was also drawn

to visualize the distribution of RAIR and RSR for these five conditions.

5.2.7.3 Correctness of Mitosis Detection (RQ3)

We used *precision* (Eq. 5.9) and *recall* (Eq. 5.10) to measure the correctness of the mitosis detection.

$$Precision = \frac{TP}{TP + FP} \tag{5.9}$$

$$Recall = \frac{TP}{TP + FN} \tag{5.10}$$

Here, we compare the precision and recall of five conditions: one-human-AI collaboration, and majority voting decisions from AI-assisted pathologists (group sizes k=3,5,7,9). Similar to the comparisons in the AI reliance metrics, for each of precision and recall, a Kruskal–Wallis and a follow-up post-hoc Dunn's test with Bonferroni correction was employed to test the significance among the condition pairs.

Because our previous work showed AI achieved higher overall performance than all participants in stage 1 [91], the "complementary team performance" in this chapter refers explicitly to cases where the human+AI approach outperforms AI (**RQ3**, *i.e.*, super-AI performance). Here, the AI operating point was selected based on the best threshold in the model validation process. For precision and recall, we defined the "success rate of achieving super-AI performance" using Equation 5.11. This equation was applied to both the one-human-AI collaboration, and majority voting decision conditions with group sizes k ranging from 3 to 27.

$$\text{Success Rate} = \frac{\text{Number of participants/runs exceeding AI performance}}{\text{Total number of participants/runs}} \times 100\% \tag{5.11}$$

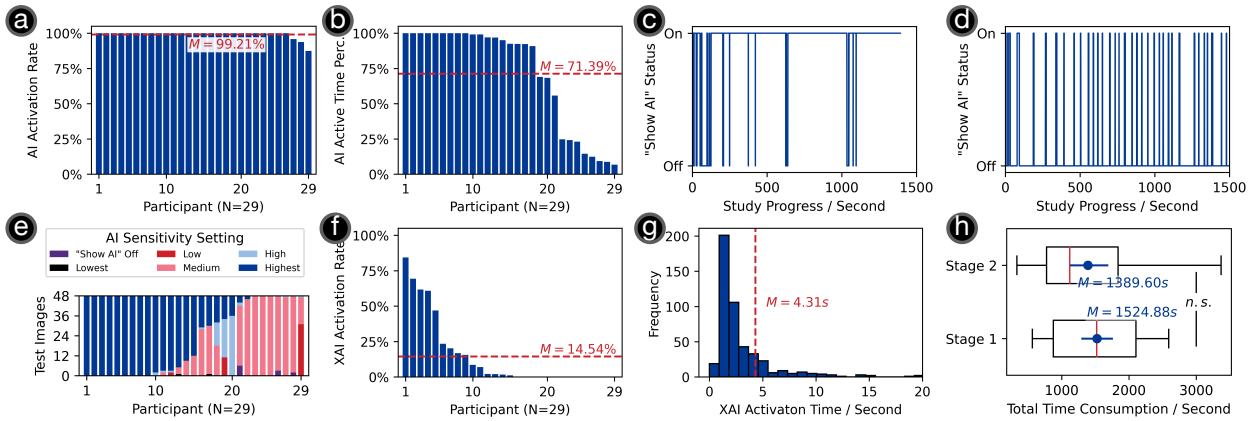


Figure 5.8: (a) Bar-plot of AI activation rates; (b) Bar-plot of AI active time percentage; Example plots showing how "Show AI" status changed for (c) a participant with a high (92.31%) AI active time percentage and (d) a participant with a low (14.48%) AI active time percentage; (e) Stacked bar-plot of participants' AI sensitivity settings; (f) XAI activation rates; (g) Histogram of XAI activation time; (h) Box-whisker plot of total time consumption of each participant spent on image examination in the stage 1 and stage 2 study. No significance (n.s.) was observed between the two stages.

5.3 Result

3/32 participants in stage 2 chose not to activate AI recommendations at all for over 45/48 test images. Therefore, they were classified as non-AI users and were excluded from subsequent analyses. For the remaining 29 participants, we report the utilization of AI and XAI in Section 5.3.1. 25/29 of the participants majorly set the sensitivity as either "highest" (N=15) or "medium" (N=10) during the stage 2 study, and they were included in the AI reliance analysis (Section 5.3.2). The responses from all 29 AI-users were used for correctness analyses in Section 5.3.3.

5.3.1 Utilization of AI and XAI

The mean AI activation rate was $M = 99.21\%$ ($SD = 0.481\%$, $CI_{95} = [98.13\%, 100.00\%]$, Figure 5.8(a))[4]. And the mean AI active time percentage was $M = 71.39\%$ ($SD = 6.713\%$, $CI_{95} = [57.75\%, 83.94\%]$, Figure 5.8(b)). 21/29 participants had $> 50\%$ AI active time percentages, with an example of how they interacted with the "Show AI" feature shown in Figure 5.8(c), which suggests the user kept the AI activated for the majority of the time, with occasional brief flickering between turning it off and on during the initial interactions. The remaining 8/29 participants had $< 25\%$ AI active time percentages: Although the "Show AI" feature was majority deactivated, these participants would still activate AI recommendations briefly while examining each image (Figure 5.8(d)). Interestingly, this pattern matches the cognitive forcing function [39] although these participants had not been instructed to do so.

For AI sensitivity settings, 15/29 participants set for the "highest" for over half of the ROI images. The remaining participants preferred to set the AI sensitivity as "high" (1/29), "medium" (10/29), "low" (1/29), or showed no clear preference (2/29), as shown in Figure 5.8(e).

Regarding XAI utilization, the mean XAI activation rate was $M = 14.54\%$ ($SD = 4.537\%$, $CI_{95} = [6.43\%, 24.25\%]$, Figure 5.8(f)). Specifically, 4/29 participants had XAI activation rates higher than 50%, while 14/29 participants did not activate any XAI at all. The mean XAI activation time was $M = 4.31$ seconds ($SD = 0.719$ seconds, $CI_{95} = [3.17 \text{ seconds}, 5.95 \text{ seconds}]$, Figure 5.8(g)).

On average, participants spent 25 minutes and 25 seconds examining all 48 test images in stage 1, and 23 minutes and 9 seconds in stage 2 (Figure 5.8(h)). The total time consumption did not show a significant difference between the two stages (Wilcoxon rank-sum test, $p = 0.31$).

[4]The mean (M), standard deviation (SD), and 95% confidence intervals (CI_{95}) were calculated by the bootstrapping method (100% re-sampling with replacement, 10,000 times)

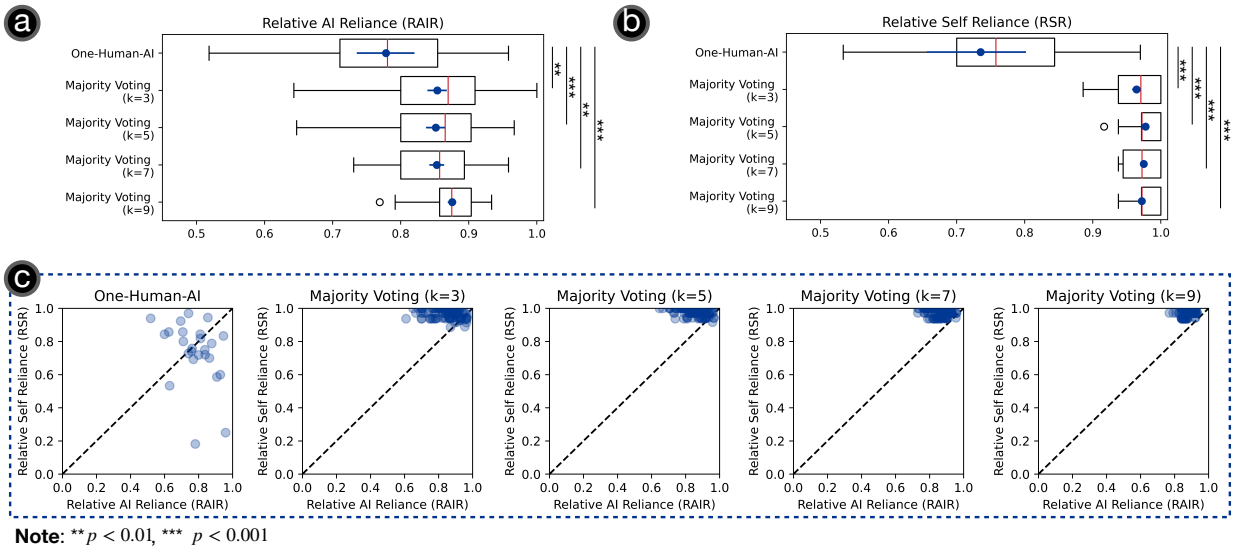


Figure 5.9: Box-whisker plots of (a) RAIR and (b) RSR for the five conditions of one-human-AI collaboration, and majority voting decisions ($k = 3, 5, 7, 9$); (c) Scatter plots for appropriateness of reliance for these five conditions.

5.3.2 Reliance on AI

As shown in Figure 5.9(a), the mean RAIR of one-human-AI collaboration was $M = 0.779$ ($SD = 0.021$, $CI_{95} = [0.735, 0.820]$). And that for majority voting decisions of group size $k = 3$ was $M = 0.852$ ($SD = 0.007$, $CI_{95} = [0.839, 0.866]$). The mean RAIR for majority voting decisions of $k = 5, 7, 9$ were 0.866, 0.861, and 0.878. All four majority voting conditions yielded higher RAIR ($\sim 9\%$ increase) than one-human-AI collaboration. A Kruskal–Wallis test showed a significant difference among the RAIR values across five conditions ($\eta^2_H = 0.043$, $p < 0.001$). Post-hoc Dunn's test with Bonferroni correction indicated significance in comparison pairs of one-human-AI *vs.*majority voting decision from group sizes of $k = 3$ ($p = 0.012$), $k = 5$ ($p < 0.001$), $k = 7$ ($p = 0.004$), and $k = 9$ ($p < 0.001$).

The mean RSR of one-human-AI collaboration was $M = 0.735$ ($SD = 0.037$, $CI_{95} = [0.657, 0.803]$, see Figure 5.9(b)). As a comparison, the mean RSR of majority voting decisions of $k = 3$ was $M = 0.964$ ($SD = 0.003$, $CI_{95} = [0.959, 0.970]$). The RSR of majority

voting decisions for $k = 5, 7, 9$ were 0.968, 0.976, and 0.967. Similarly, all four majority voting conditions led to higher RSR ($\sim$ 31% increase). A Kruskal–Wallis test showed a significant difference among the RSR values across five conditions ($\eta^2_H = 0.178$, $p < 0.001$). Post-hoc Dunn-Bonferroni test showed significance in comparison pairs of one-human-AI *vs.*majority voting decision from group sizes of $k = 3$ ($p < 0.001$), $k = 5$ ($p < 0.001$), $k = 7$ ($p < 0.001$), and $k = 9$ ($p < 0.001$).

Figure 5.9(c) presents the appropriateness of reliance (AoR) scatter plots for five conditions. These plots demonstrate that majority voting decisions could improve higher RAIR and RSR simultaneously, indicating a high level of AoR was achieved.

5.3.3 Correctness of Mitosis Detection

As shown in Figure 5.10(a), the mean precision of one-human-AI collaboration was $M = 0.824$ ($SD = 0.023$, $CI_{95} = [0.776, 0.867]$). For majority voting decisions of $k = 3$, the mean precision was $M = 0.902$ ($SD = 0.004$, $CI_{95} = [0.893, 0.910]$). The majority voting of $k = 5, 7, 9$ had mean precisions of 0.924, 0.934, and 0.934, respectively. The AI achieved a higher precision of 0.961. All four majority voting conditions achieved higher precision ($\sim$ 8% increase) than the one-human-AI collaboration. A Kruskal–Wallis test showed that the precision significantly differed across five conditions ($\eta^2_H = 0.150$, $p < 0.001$). Post-hoc Dunn-Bonferroni test did not observe significance in the comparison pair of one-human-AI *vs.*majority voting decision $k = 3$ ($p = 0.715$). Statistical significance was observed for comparison pairs of one-human-AI *vs.*majority voting decision $k = 5$ ($p < 0.001$), $k = 7$ ($p < 0.001$), and $k = 9$ ($p < 0.001$).

The mean recall of one-human-AI collaboration was $M = 0.817$ ($SD = 0.013$, $CI_{95} = [0.790, 0.841]$, see Figure 5.10(b)). Majority voting decisions of $k = 3$ had a mean recall of $M = 0.843$ ($SD = 0.003$, $CI_{95} = [0.838, 0.851]$). Majority voting decisions of $k = 5, 7, 9$ had mean precisions of 0.850, 0.851, and 0.850. In comparison, AI achieved a precision of 0.841. Kruskal–Wallis test did not show that the recall differed significantly across five conditions

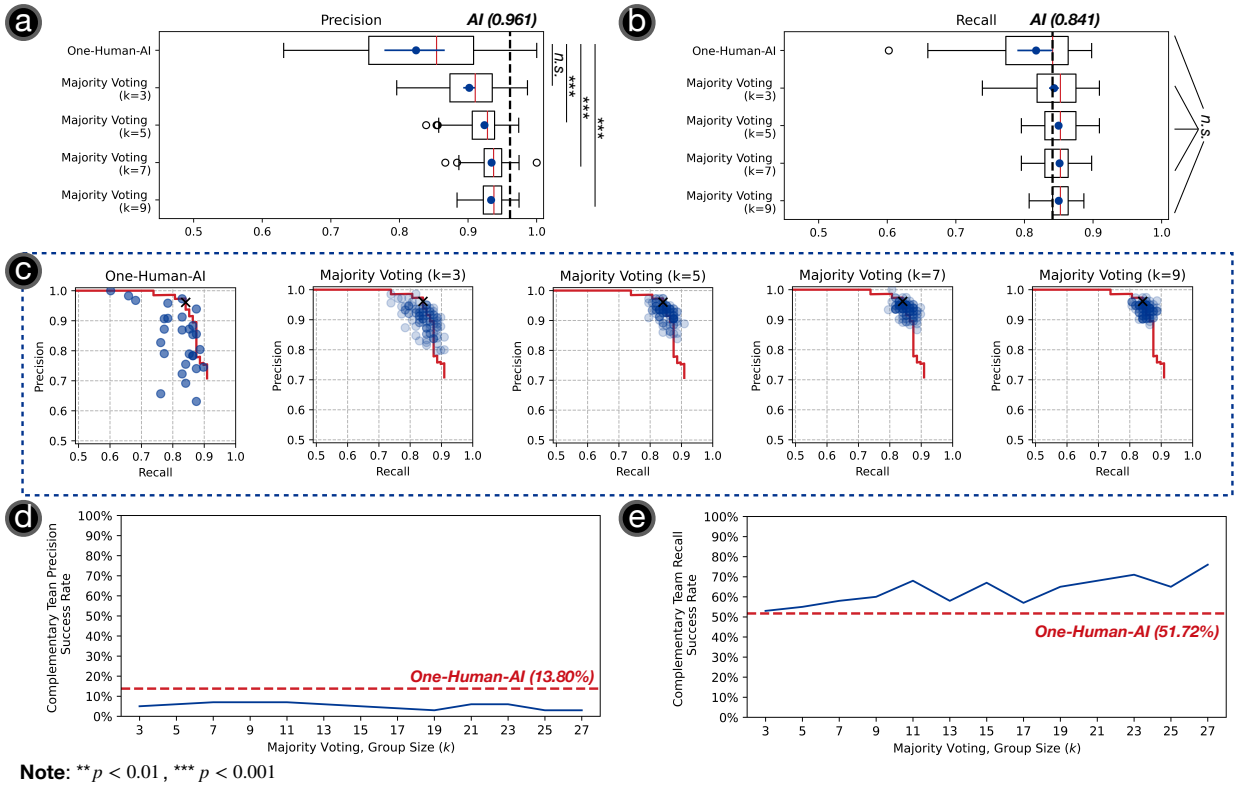


Figure 5.10: Box-whisker plots of precision and recall for the five conditions of one-human-AI collaboration, and majority voting decisions ($k = 3, 5, 7, 9$); (c) Precision-recall plots for mitosis detection for these five conditions. The red line represents the precision-recall curve of AI, and the 'x' marker indicates the AI's performance at a threshold determined by the best validation performance. The success rates of achieving super-AI performance (*i.e.*, percentage of human+AI cases where their performance is higher than both humans and AI) for the criteria of (d) precision and (e) recall.

($\eta_H^2 < 0.001$, $p = 0.774$).

Figure 5.10(c) presents the precision-recall scatter plots for the five conditions. The plots reveal that the majority voting decision exhibits lower variation in both precision and recall compared to the one-human-AI collaboration, indicating a more robust performance. This observation is further supported by the lower SD values for the majority voting decisions, as reported above.

Regarding the success rates for achieving super-AI performance, for precision, none of the majority voting conditions (*i.e.*, $k = 3 \rightarrow 27$) was higher than the success rate achieved by one-human-AI collaboration (13.87% success rate, Figure 5.10(d)). On the other hand, for recall, all majority voting conditions had higher success rates compared to one-human-AI collaboration (51.72% success rate): As shown in Figure 5.10(e), the lowest success rate was observed at $k = 3$ (53% success rate), and the highest was achieved at $k = 27$, reaching 76%.

5.4 Discussion

5.4.1 Summary of Result

5.4.1.1 Summary of RQ1

For most participants, AI was activated at least once in most images. However, this does not imply that the AI was constantly active throughout the entire study. Notably, 8/29 participants deactivated AI for most of the study, and only activated it briefly occasionally. That is, in certain instances, the 'AI on-request' feature posed cognitive forcing function effects.

The utilization of XAI was relatively low; only four participants opened more than 50% of the XAI evidence, while nearly half of the participants did not open any. Even when XAI was opened, the time spent by participants on viewing XAI was relatively short (about four seconds) – in the context of pathologist-AI collaboration, the effectiveness of XAI in

mitigating over-reliance may be limited. This is likely because the time-pressing nature of the pathology task outweighed the benefit of XAI explanations, causing pathologists to use XAI less in practice. In light of this, we argue that alternative approaches, such as the majority voting used in this study, need to be investigated to enable appropriate AI reliance for future pathology applications.

5.4.1.2 Summary of RQ2

Pair-wise statistical tests revealed significant improvements in both RAIR and RSR metrics for majority voting decisions ($k = 3, 5, 7, 9$), compared to one pathologist collaborating with AI. Specifically, RAIR showed an approximate 9% increase, and RSR showed about 31% increase. The PAIR-RSR scatter plots indicated simultaneous improvements in both metrics. Such results demonstrate a reduction in the proportion of over-reliance against correct self-reliance events, and under-reliance against correct AI reliance events, indicating a higher level of appropriateness of reliance was achieved (according to the definitions in [180]).

5.4.1.3 Summary of RQ3

No significant difference in the precision was observed between the condition of one-human-AI collaboration and the majority voting with $k = 3$. A statistical significance in the precision was observed when increasing k to 5, 7, and 9. The majority voting conditions improved precision by approximately 8%. For recall, no significant differences were observed. The precision-recall scatter plots demonstrated that majority voting decisions exhibited lower variation, suggesting that they were robust and less prone to be influenced by the sample selection.

All majority voting conditions for $k = 3 \rightarrow 27$ did not show a higher success rate in achieving super-AI precision than one-human-AI collaboration. This is because AI had a

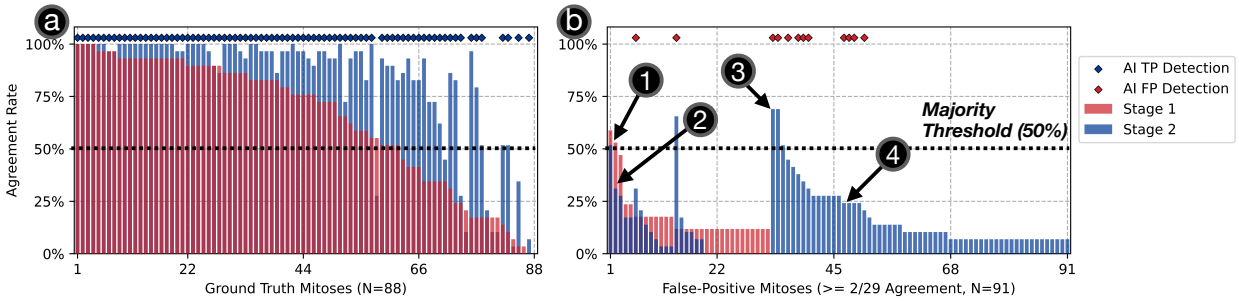


Figure 5.11: Bar plots for the agreements among 29 participants for (a) 88 ground-truth mitoses, and (b) 91 false-positive mitoses that at least two participants agreed on. The diamond markers ($\Diamond$) stand for the AI detections under the "Highest" AI sensitivity setting. ① An example of under-reliance that might not be addressed by the majority voting; ② An example of under-reliance that might be addressed by the majority voting; ③ An example of over-reliance that might not be addressed by the majority voting; and ④ An example of over-reliance that might be addressed by the majority voting.

high precision of 0.961, and there was a lack of space for improvement. For recall, all majority voting conditions ($k = 3 \rightarrow 27$) showed higher success rates. Notably, the highest success rate, 76%, was achieved at $k = 27$, indicating a 46.95% increase over the one-human-AI collaboration condition (51.72% success rate).

5.4.2 The Mechanism and Cost of Majority Voting

To further explore why the majority voting mechanism was effective, we introduced a metric, "agreement rate," defined as the percentage of participants the reported a cell as a mitosis (regardless of its actual status). We calculated the agreement rates of the 29 participants in both stage 1 and stage 2 studies. These agreement rates covered all 88 ground truth mitoses (Figure 5.11(a)) and 91 false-positive mitoses reported by at least two participants (Figure 5.11(b)). According to Section 5.2.6, cells with agreement rates higher than 50% should be kept as the majority voting decisions. While Figure 5.11 is not directly applicable for

interpreting results in smaller sub-groups (*e.g.*, $k = 3$), it illustrates the general trends in participants' agreement rates when influenced by AI. The data revealed two key insights:

- **Reducing Over-Reliance on AI False Positives**: AI's false-positive detections led to higher agreement rates among participants (as shown in Figure 5.11(b)③), suggesting participants' tendency of over-reliance in at stage 2. The majority of these false-positive detections did not achieve agreement rates higher than 50% (Figure 5.11(b)④). In other words, from a group's perspective, it was not usual for the majority of participants to consistently over-rely when AI made false-positive mistakes. Therefore, the over-reliance can be reduced by the majority voting mechanism.

- **Reducing Under-Reliance on Human False Positives:** At stage 2, participants may make the same false-positive mistake as in stage 1, even when AI correctly suggested negative (Figure 5.11(b)①). This suggests that the under-reliance incidents happened when one participant collaborated with AI. Nevertheless, agreement rates for these false positives rarely exceeded the 50% majority threshold (Figure 5.11(b)②), indicating that majority voting could reduce under-reliance.

To understand the underlying cost of the majority voting mechanism, we analyzed time consumption spent on employing multiple pathologists, and its association with the correctness. Specifically, we conducted 100 majority voting runs for each group size (k) ranging from 3 to 27. We applied Pearson's correlation analysis to assess the relationship between precision or recall achieved in each run and its corresponding time consumption. We found a moderate positive correlation between precision and time consumption (Pearson's $r = 0.39$, $p < 0.001$, $N = 1,300$, Figure 5.12(a)), and a weak positive correlation between recall and time consumption (Pearson's $r = 0.14$, $p < 0.001$, $N = 1,300$, Figure 5.12(b)). Certain runs with a relatively small time consumption could reach considerable precision and recall. Note that this is a 'bare minimum' estimation: Delays caused by coordinating pathologists should be taken into account in practical applications.

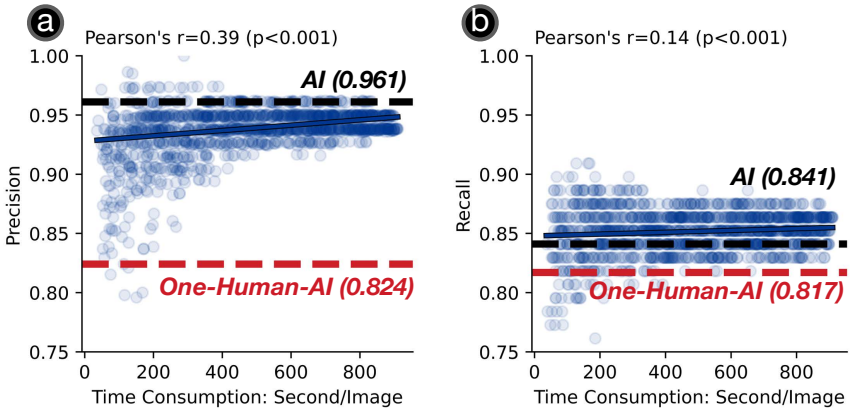


Figure 5.12: Linear regression plots studying the relations between (a) precision-time consumption, and (b) recall-time consumption while synthesizing majority voting decisions, $k = 3 \rightarrow 27$, $n = 100$ for each k.

5.4.3 On Developing Structured Decision-Making Processes with AI+k

Different from traditional one-human-AI collaboration (AI+1), this study sets the first step towards multiple medical professionals collaborating with AI (AI+k) using a simple majority voting technique. We argue that this majority voting approach has three advantages: *(i)* It is flexible and has a simple structure, eliminating the need for face-to-face or online discussions; *(ii)* It keeps participants anonymous, thus reducing potential social pressure; *(iii)* It is inherently democratic, ensuring that each participant's opinion has an equal weight. We found that this majority voting approach could effectively improve the appropriateness of reliance, and achieve higher-quality medical decisions. As for the limitations of majority voting, one may argue that this approach does not incorporate the discussion process, and decisions with conflicts (*i.e.*, $\sim 50\%$ agreement rates) cannot be addressed easily.

Future works might explore AI+k decision-making techniques that involve structured or semi-structured face-to-face discussions [32]. Traditionally, these discussions were moderated by the humans. Nonetheless, we envision that future AI can not only help each group member to reach a decision (*e.g.*, help pathologists detect mitoses in this study), but can moderate

the discussions. For instance, a large language model (LLM) [148] might anonymously gather and summarize comments from each group member and present a consolidated overview to the group. Members could then have an opportunity to revise their decisions after hearing from the LLM's summary. Given the LLM's omni-availability, no conflict of interest, and and impartiality to authority or personal factors, such AI-facilitated discussions could offer advantages in speed and bias correction, compared to traditional discussion coordination with human moderators.

5.4.4 Towards Efficient and Reliable Medical Decisions with AI+k

Section 5.4.2 showed that the performance of majority voting decisions from AI+k showed a positive relation to the time consumption. In other words, in general, the more medical professionals involved, the higher the quality of the majority voting decision. Typically, high-risk medical decisions involve 7–10 group members [145], while groups as large as 27 done in this study were quite rare. Therefore, considering the time taken to reach a result, we argue that not all medical decisions necessitate the AI+{*large* k} approach: cases with high confidence from both AI and humans could be adjudicated by smaller groups with as few as three experts, while those with low AI confidence or prone to human errors could benefit from incrementally larger group sizes, which can yield better and more robust outcomes.

Determining the optimal balance between decision-making and time expenditure has been well-explored in previous crowd-sourcing works [63]. However, one should be aware that the workflow of medical professionals is usually different from that of general users, and their preferences in using AI and XAI may also vary (as shown in Section 5.3.1). Therefore, future research should focus on exploring which AI+k methods can seamlessly integrate into the workflow of medical professionals, effectively balancing efficiency and reliability in the medical decisions of multiple doctors. Additionally, investigating the role of counterfactual explanations [229, 64] to build trust and facilitate appropriate AI reliance could complement approaches like majority voting, potentially improving interpretability and familiarity with

the decision process when integrating AI into risk-sensitive medical workflows.

5.4.5 Limitations and Future Work

The following points are the limitations of this study and are regarded as future work.

- The majority voting synthesizing process did not involve any discussion or communication among participants, which could influence the outcomes.
- A 50% threshold was used to represent the majority. Other thresholds and their impacts were not investigated.
- The potential learning effect, particularly among participants in training (*i.e.*, residents and medical students), between stage 1 and stage 2 of the study cannot be ignored.
- All participants were from one country, potentially limiting the generalizability of findings.

5.5 Conclusion

This chapter introduces and validates the majority voting approach to enable doctors' appropriate reliance on medical AI. By recruiting 32 pathology professionals, we conducted a multi-institutional, multi-stage user study focusing on detecting mitoses in tumor images. Our analysis revealed that even with groups of three doctors, the majority-voting decisions had a higher appropriateness of AI reliance, compared to one doctor collaborating with AI. Subsequently, the majority voting decisions demonstrated increased precision and recall, although no statistical significance in recall was observed. Additionally, majority voting decisions were more likely to achieve super-AI performance in the recall. While effective on its own, majority voting can also be used together with other techniques to enable appropriate AI reliance. Involving multiple experts in decision-making can yield higher-quality, more

robust outcomes that are less prone to AI errors, which holds promise in pathology and broader high-stakes domains.

CHAPTER 6

Summary

This thesis explores the landscape of human-AI collaborative ecosystems in pathology through both qualitative and quantitative investigations. Through a multi-faceted approach – including field studies, artifact development and validation, and empirical evaluations – it examines how AI can effectively support pathologists in clinical decision-making. The overarching goal is to develop a pathology assistant that enhances *correctness*, *efficiency*, and *safety* in high-stakes diagnostic workflows.

This chapter first summarizes the answers to the three key research questions and findings, followed by an outlook on the future of AI-assisted pathology environments.

6.1 RQ1: How should human-AI collaboration systems be designed for pathology, and how can these insights inform future system development?

Chapter 2 outlined six key recommendations for designing effective human-AI collaboration systems in pathology, which can be further distilled into three overarching principles:

1. AI assistance should be *spatially comprehensive* across multiple magnifications, *temporally continuous*, and provide guidance that is *simple*, *intuitive*, and *actionable* for pathologists.

2. Human-AI collaborative workflows should *align with pathologists' existing ones* to min-

imize learning costs, optimizing *challenging step*s while *preserving their decision autonomy.*

3. If introducing new workflow components is necessary, their benefits *must outweigh the additional effort and time* required from pathologists.

Building on these principles, Chapter 3 identified three key observations from pathologists' navigation preferences: *(i)* overview first, then detail, *(ii)* using macroscopic patterns to locate ROIs in the low magnification, and *(iii)* low throughput in higher magnifications. Reflecting these findings, NaviPath was designed with three unique features: *(i)* Hierarchical AI Recommendations, *(ii)* Customizable Recommendations by Multiple Criteria, and *(iii)* Cue-Based Navigation for High Magnifications.

Furthermore, Chapter 4 addressed the challenges of *comprehensiveness*, *explainability*, and *integrability* in AI-assisted complex pathology diagnosis tasks by introducing two key designs in xPath: *(i)* Joint-Analyses of Multiple Criteria, and *(ii)* Explanation by Hierarchically Traceable Evidence for Each Criterion.

Finally, Chapter 5, adopted the majority voting to enhance the reliability of human-AI collaborative decision-making – an approach already commonly used by pathologists when signing out challenging cases.

The efficacy of these designs was further validated through user studies with pathologists and trainees, with key results summarized in the next section.

6.2 RQ2: How does human-AI collaboration affect pathologists' examination and diagnostic processes?

The key performance indicators for evaluating human-AI workflows include the following aspects:

1. **Efficiency** is measured by the number of target pathology patterns examined per unit time. In Chapter 3, NaviPath enabled participants to identify more than twice the number of mitoses per unit time compared to the manual system. Although Chapter 4 did not measure efficiency, participants reported significantly higher comprehensiveness in their examination processes, indirectly suggesting improved efficiency. It is important to the difference between *efficiency* and *time consumption*; in all three studies in Chapters 3, 4, and 5, no significant reduction in examination time was observed.

2. **Correctness** is measured with accuracy, precision, and recall (sensitivity). In Chapter 3, users achieved higher precision and recall in mitosis detection using NaviPath compared to both manual examination or standalone AI. In Chapter 4, diagnoses made using xPath demonstrated had higher accuracy than those made with manual examination. In Chapter 5, majority voting of even three AI-assisted pathologists significantly improved precision, though recall showed less significance compared to conventional AI-assisted approaches.

3. **Reliability of Decision Outcomes** is measured through the appropriateness of AI reliance, measured by the proportion of over-reliance and under-reliance incidents. In Chapter 5, the majority voting of three AI-assisted pathologists significantly increased Relative AI Reliance and Relative Self-Reliance, indicating a reduction in both over-reliance and under-reliance events. Additionally, the variance in majority-voted decisions was significantly reduced, indicating more reliable decision outcomes.

4. **Perceived Acceptance** is measured with user self-reported Likert-scale measures through surveys, including workload, confidence, and system preference. In both Chapters 3 and 4, a significant reduction in workload was observed, alongside a strong user preference for AI-assisted workflows over manual examination. Additionally, users reported high confidence in their decisions when using AI-assisted systems.

6.3 RQ3: How can human-AI collaboration be optimized to maximize pathologists' correctness while ensuring appropriate AI reliance?

Chapter 5 reported that XAI was rarely referenced in AI-assisted decision-making, thus having limited effect in regulating inappropriate AI reliance incidents. To overcome this limitation, this chapter introduces a majority voting mechanism by ensembling decisions from three or more odd numbers of AI-assisted pathologists. Through a nationwide, six-month-long survey with 32 pathologists and trainees, the study demonstrates that majority voting with three or more AI-assisted pathologists significantly reduces both under-reliance and over-reliance incidents, resulting in more appropriate decision outcomes. An interesting finding from this chapter is that, in inappropriate reliance incidents, participants' agreement rates rarely exceeded 50%. This finding makes majority voting an effective strategy to mitigate inappropriate AI reliance by preventing a single-point failure from compromising the overall decision outcome.

6.4 Concluding Remarks: A Future of Digital Pathology with Omni-Available, 24/7 AI

With its high throughput and cost-effectiveness, AI will make large-cohort, multi-center retrospective analyses more available, which will expand the possibilities for data-driven pathology research. In the future, more studies will quantitatively evaluate the interrelationships between histopathological features and prognosis, which will shed light on new computational pathology-informed medical standards. These advancements will provide quantitative evidence for the making of medical guidelines and generate meaningful pathognomonic insights. However, in clinical practice, AI will continue to function as a "software as a medical device", which will offer comprehensive *assistance* across various pathology tasks

while enabling pathologists to retain ultimate decision-making autonomy over AI-generated recommendations. Moving forward, a pathologist-centered approach will be preferred in designing, developing, and studying interactive computational pathology for clinical workflows, incorporating pathologists' expertise and enhancing their diagnostic capabilities.

For instance, AI assistants could be developed for a dynamic, multimodal pathology environment, where a pathologist-AI collaboration interface integrates H&E, IHC, FISH, and NGS data according to medical guidelines. Such a system could suggest follow-up tests based on initial findings, continuously refine risk assessments using fine-tuned foundation models, and generate preliminary reports via LLMs for pathologists to review. Such systems could serve not only as a clinical tool but also as a valuable resource for training and education.

Additionally, pathology decision-making may further benefit from group intelligence. Future research could explore more efficient pathways for fostering high-quality diagnostic decisions with minimal personnel, such as whether two collaborating pathologists could achieve better diagnostic outcomes than working independently. Could one pathologist oversee another's work? Could structured discussions before and after decision-making facilitate consensus development?

Finally, the future of digital pathology may see the application of virtual pathologists available on call 7/24: Second-opinion consultations are a critical component of complex case evaluations, yet accessing an available pathologist around the clock remains challenging. Could AI chatbots serve as an alternative for second-opinion consultations? Future studies could explore how traditional inter-pathologist communication paradigms might be reimagined by leveraging real-world pathologist conversations. A pathology vision-language model could be fine-tuned to emulate human pathologists' discourse, addressing issues related to consultation delays and the biases of the geographic distribution of specialists.

REFERENCES

[1] Jacob T Abel, Peter Ouillette, Christopher L Williams, John Blau, Jerome Cheng, Keluo Yao, Winston Y Lee, Toby C Cornish, Ulysses GJ Balis, and David S McClintock. Display characteristics and their impact on digital pathology: A current review of pathologists' future "microscope". *Journal of Pathology Informatics*, 11(1):23, 2020.

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

[3] Maximilian Alber, Stephan Tietz, Jonas Dippel, Timo Milbich, Timothée Lesort, Panos Korfiatis, Moritz Krügener, Beatriz Perez Cancer, Neelay Shah, Alexander Möllers, et al. A novel pathology foundation model by mayo clinic, charit\'e, and aignostics. *arXiv preprint arXiv:2501.05409*, 2025.

[4] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, et al. Guidelines for human-ai interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, page 3. ACM, 2019.

[5] Mohamed Amgad, Lamees A Atteya, Hagar Hussein, Kareem Hosny Mohammed, Ehab Hafiz, Maha AT Elsebaie, Ahmed M Alhusseiny, Mohamed Atef AlMoslemany, Abdelmagid M Elmatboly, Philip A Pappalardo, et al. Nucls: A scalable crowdsourcing, deep learning approach and dataset for nucleus classification, localization and segmentation. *arXiv preprint arXiv:2102.09099*, 2021.

[6] 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. Structured crowdsourcing enables convolutional segmentation of histology images. *Bioinformatics*, 35(18):3461–3467, 2019.

[7] Sophia K Apple. Sentinel lymph node in breast cancer: review article from a pathologist's point of view. *Journal of pathology and translational medicine*, 50(2):83, 2016.

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

[9] 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. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58:82–115, 2020.

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

[11] Marc Aubreville, Christof Bertram, Mitko Veta, Robert Klopfleisch, Nikolas Stathonikos, Katharina Breininger, Natalie ter Hoeve, Francesco Ciompi, and Andreas Maier. Quantifying the scanner-induced domain gap in mitosis detection. *arXiv preprint arXiv:2103.16515*, 2021.

[12] Marc Aubreville, Christof A. Bertram, Taryn A. Donovan, Christian Marzahl, Andreas Maier, and Robert Klopfleisch. A completely annotated whole slide image dataset of canine breast cancer to aid human breast cancer research. *Scientific Data*, 7(1):417, November 2020.

[13] Marc Aubreville, Nikolas Stathonikos, Christof A. Bertram, Robert Klopfleisch, Natalie ter Hoeve, Francesco Ciompi, Frauke Wilm, Christian Marzahl, Taryn A. Donovan, Andreas Maier, Jack Breen, Nishant Ravikumar, Youjin Chung, Jinah Park, Ramin Nateghi, Fattaneh Pourakpour, Rutger H.J. Fick, Saima Ben Hadj, Mostafa Jahanifar, Adam Shephard, Jakob Dexl, Thomas Wittenberg, Satoshi Kondo, Maxime W. Lafarge, Viktor H. Koelzer, Jingtang Liang, Yubo Wang, Xi Long, Jingxin Liu, Salar Razavi, April Khademi, Sen Yang, Xiyue Wang, Ramona Erber, Andrea Klang, Karoline Lipnik, Pompei Bolfa, Michael J. Dark, Gabriel Wasinger, Mitko Veta, and Katharina Breininger. Mitosis domain generalization in histopathology images — the midog challenge. *Medical Image Analysis*, 84:102699, 2023.

[14] Marc Aubreville, Nikolas Stathonikos, Taryn A. Donovan, Robert Klopfleisch, Jonas Ammeling, Jonathan Ganz, Frauke Wilm, Mitko Veta, Samir Jabari, Markus Eckstein, Jonas Annuscheit, Christian Krumnow, Engin Bozaba, Sercan Çayır, Hongyan Gu, Xiang 'Anthony' Chen, Mostafa Jahanifar, Adam Shephard, Satoshi Kondo, Satoshi Kasai, Sujatha Kotte, V.G. Saipradeep, Maxime W. Lafarge, Viktor H. Koelzer, Ziyue Wang, Yongbing Zhang, Sen Yang, Xiyue Wang, Katharina Breininger, and Christof A. Bertram. Domain generalization across tumor types, laboratories, and species — insights from the 2022 edition of the mitosis domain generalization challenge. *Medical Image Analysis*, 94:103155, 2024.

[15] Murat Seckin Ayhan and Philipp Berens. Test-time data augmentation for estimation of heteroscedastic aleatoric uncertainty in deep neural networks. 2018.

[16] Wei Ba, Shuhao Wang, Meixia Shang, Ziyan Zhang, Huan Wu, Chunkai Yu, Ranran Xing, Wenjuan Wang, Lang Wang, Cancheng Liu, et al. Assessment of deep learning assistance for the pathological diagnosis of gastric cancer. *Modern Pathology*, 35(9):1262–1268, 2022.

[17] Boris Babenko. Multiple instance learning: algorithms and applications. *View Article PubMed/NCBI Google Scholar*, pages 1–19, 2008.

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

[19] Maschenka CA Balkenhol, David Tellez, Willem Vreuls, Pieter C Clahsen, Hans Pinckaers, Francesco Ciompi, Peter Bult, and Jeroen AWM van der Laak. Deep learning assisted mitotic counting for breast cancer. *Laboratory investigation*, 99(11):1596–1606, 2019.

[20] Peter Bankhead, Maurice B Loughrey, José A Fernández, Yvonne Dombrowski, Darragh G McArt, Philip D Dunne, Stephen McQuaid, Ronan T Gray, Liam J Murray, Helen G Coleman, et al. Qupath: Open source software for digital pathology image analysis. *Scientific reports*, 7(1):16878, 2017.

[21] Gagan Bansal, Besmira Nushi, Ece Kamar, Walter S Lasecki, Daniel S Weld, and Eric Horvitz. Beyond accuracy: The role of mental models in human-ai team performance. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 7, pages 2–11, 2019.

[22] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA, 2021. Association for Computing Machinery.

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

[24] Patrick Baudisch and Ruth Rosenholtz. Halo: A technique for visualizing off-screen objects. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '03, page 481–488, New York, NY, USA, 2003. Association for Computing Machinery.

[25] Benjamin B Bederson, James D Hollan, Ken Perlin, Jonathan Meyer, David Bacon, and George Furnas. Pad++: A zoomable graphical sketchpad for exploring alternate interface physics. *Journal of Visual Languages & Computing*, 7(1):3–32, 1996.

[26] Emma Beede, Elizabeth Baylor, Fred Hersch, Anna Iurchenko, Lauren Wilcox, Paisan Ruamviboonsuk, and Laura M. Vardoulakis. 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*, CHI '20, page 1–12, New York, NY, USA, 2020. Association for Computing Machinery.

[27] 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. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *Jama*, 318(22):2199–2210, 2017.

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

[29] Christof A. Bertram, Marc Aubreville, Corinne Gurtner, Alexander Bartel, Sarah M. Corner, Martina Dettwiler, Olivia Kershaw, Erica L. Noland, Anja Schmidt, Dodd G. Sledge, Rebecca C. Smedley, Tuddow Thaiwong, Matti Kiupel, Andreas Maier, and Robert Klopfleisch. Computerized calculation of mitotic count distribution in canine cutaneous mast cell tumor sections: Mitotic count is area dependent. *Veterinary Pathology*, 57(2):214–226, 2020.

[30] Christof A Bertram, Marc Aubreville, Christian Marzahl, Andreas Maier, and Robert Klopfleisch. A large-scale dataset for mitotic figure assessment on whole slide images of canine cutaneous mast cell tumor. *Scientific data*, 6(1):1–9, 2019.

[31] Wenya Linda Bi, Ahmed Hosny, Matthew B Schabath, Maryellen L Giger, Nicolai J Birkbak, Alireza Mehrtash, Tavis Allison, Omar Arnaout, Christopher Abbosh, Ian F Dunn, et al. Artificial intelligence in cancer imaging: clinical challenges and applications. *CA: a cancer journal for clinicians*, 69(2):127–157, 2019.

[32] Nick Black, Maggie Murphy, Donna Lamping, Martin McKee, Colin Sanderson, Janet Askham, and Theresa Marteau. Consensus development methods: a review of best practice in creating clinical guidelines. *Journal of health services research & policy*, 4(4):236–248, 1999.

[33] Marsden S. Blois. Clinical judgment and computers. *New England Journal of Medicine*, 303(4):192–197, 1980. PMID: 7383090.

[34] Felix Bork, Christian Schnelzer, Ulrich Eck, and Nassir Navab. Towards efficient visual guidance in limited field-of-view head-mounted displays. *IEEE Transactions on Visualization and Computer Graphics*, 24(11):2983–2992, 2018.

[35] Kenza Bouzid, Harshita Sharma, Sarah Killcoyne, Daniel C Castro, Anton Schwaighofer, Max Ilse, Valentina Salvatelli, Ozan Oktay, Sumanth Murthy, Lucas Bordeaux, et al. Enabling large-scale screening of barrett's esophagus using weakly supervised deep learning in histopathology. *Nature Communications*, 15(1):2026, 2024.

[36] 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. 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):993–1007, 2008.

[37] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[38] Christine Bruce, Ioannis Prassas, Mark Mokhtar, Blaise Clarke, Elaria Youssef, Catherine Wang, and George M Yousef. Transforming diagnostics: The implementation of digital pathology in clinical laboratories. *Histopathology*, 2024.

[39] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. To trust or to think: Cognitive forcing functions can reduce overreliance on ai in ai-assisted decision-making. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW1), apr 2021.

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

[41] Wouter Bulten, Kimmo Kartasalo, Po-Hsuan Cameron Chen, Peter Ström, Hans Pinckaers, Kunal Nagpal, Yuannan Cai, David F. Steiner, Hester van Boven, Robert Vink, Christina Hulsbergen-van de Kaa, Jeroen van der Laak, Mahul B. Amin, Andrew J. Evans, Theodorus van der Kwast, Robert Allan, Peter A. Humphrey, Henrik Grönberg, Hemamali Samaratunga, Brett Delahunt, Toyonori Tsuzuki, Tomi Häkkinen, Lars Egevad, Maggie Demkin, Sohier Dane, Fraser Tan, Masi Valkonen, Greg S. Corrado, Lily Peng, Craig H. Mermel, Pekka Ruusuvuori, Geert Litjens, Martin Eklund, Américo Brilhante, Aslı Çakır, Xavier Farré, Katerina Geronatsiou, Vincent Molinié, Guilherme Pereira, Paromita Roy, Günter Saile, Paulo G. O. Salles, Ewout Schaafsma, Joëlle Tschui, Jorge Billoch-Lima, Emíio M. Pereira, Ming Zhou, Shujun He, Sejun Song, Qing Sun, Hiroshi Yoshihara, Taiki Yamaguchi, Kosaku Ono, Tao Shen, Jianyi Ji, Arnaud Roussel, Kairong Zhou, Tianrui Chai, Nina Weng, Dmitry Grechka, Maxim V. Shugaev, Raphael Kiminya, Vassili Kovalev, Dmitry Voynov, Valery Malyshev, Elizabeth Lapo, Manuel Campos, Noriaki Ota, Shinsuke Yamaoka, Yusuke Fujimoto, Kentaro Yoshioka, Joni Juvonen, Mikko Tukiainen, Antti Karlsson, Rui Guo, Chia-Lun Hsieh, Igor Zubarev, Habib S. T. Bukhar, Wenyuan Li, Jiayun Li, William Speier, Corey Arnold, Kyungdoc Kim, Byeonguk Bae, Yeong Won Kim, Hong-Seok Lee, Jeonghyuk Park, and the PANDA challenge consortium. Artificial intelligence for diagnosis and Gleason grading of prostate cancer: the PANDA challenge. *Nature Medicine*, 28(1):154–163, January 2022.

[42] Adrian Bussone, Simone Stumpf, and Dympna O'Sullivan. The role of explanations on trust and reliance in clinical decision support systems. In *2015 International Conference on Healthcare Informatics*, pages 160–169, 2015.

[43] Carrie J Cai, Emily Reif, Narayan Hegde, Jason Hipp, Been Kim, Daniel Smilkov, Martin Wattenberg, Fernanda Viegas, Greg S Corrado, Martin C Stumpe, et al. 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*, pages 1–14, 2019.

[44] Carrie J. Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. "hello ai": Uncovering the onboarding needs of medical practitioners for human-ai collaborative decision-making. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW), nov 2019.

[45] Francisco M. Calisto, Alfredo Ferreira, Jacinto C. Nascimento, and Daniel Gonçalves. Towards touch-based medical image diagnosis annotation. In *Proceedings of the 2017 ACM International Conference on Interactive Surfaces and Spaces*, ISS '17, page 390–395, New York, NY, USA, 2017. Association for Computing Machinery.

[46] Francisco Maria Calisto, Nuno Nunes, and Jacinto C. Nascimento. Breastscreening: On the use of multi-modality in medical imaging diagnosis. In *Proceedings of the International Conference on Advanced Visual Interfaces*, AVI '20, New York, NY, USA, 2020. Association for Computing Machinery.

[47] Francisco Maria Calisto, Carlos Santiago, Nuno Nunes, and Jacinto C. Nascimento. Introduction of human-centric ai assistant to aid radiologists for multimodal breast image classification. *International Journal of Human-Computer Studies*, 150:102607, 2021.

[48] Francisco Maria Calisto, Carlos Santiago, Nuno Nunes, and Jacinto C. Nascimento. Breastscreening-ai: Evaluating medical intelligent agents for human-ai interactions. *Artificial Intelligence in Medicine*, 127:102285, 2022.

[49] Shiye Cao and Chien-Ming Huang. Understanding user reliance on ai in assisted decision-making. *Proc. ACM Hum.-Comput. Interact.*, 6(CSCW2), nov 2022.

[50] Wenzhi Cao, Vahid Mirjalili, and Sebastian Raschka. Rank consistent ordinal regression for neural networks with application to age estimation. *Pattern Recognition Letters*, 140:325–331, 2020.

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

[52] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 839–847, 2018.

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

[54] Richard J Chen, Tong Ding, Ming Y Lu, Drew FK Williamson, Guillaume Jaume, Andrew H Song, Bowen Chen, Andrew Zhang, Daniel Shao, Muhammad Shaban, et al. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 30(3):850–862, 2024.

[55] Jinwook Choi, Kyoungbun Lee, Won-Ki Jeong, and Se Young Chun. Paip2021: Perineural invasion in multiple organ cancer (colon, prostate, and pancreatobiliary tract), Mar 2021.

[56] Dan C Cireşan, Alessandro Giusti, Luca M Gambardella, and Jürgen Schmidhuber. Mitosis detection in breast cancer histology images with deep neural networks. In *International conference on medical image computing and computer-assisted intervention*, pages 411–418. Springer, 2013.

[57] Andy Cockburn, Amy Karlson, and Benjamin B Bederson. A review of overview+ detail, zooming, and focus+ context interfaces. *ACM Computing Surveys (CSUR)*, 41(1):1–31, 2009.

[58] Y.U.I. Collan, T. Kuopio, J.P.A. Baak, R. Becker, W.V. Bogomoletz, M. Deverell, P. Van Diest, C. Van Galen, K. Gilchrist, A. Javed, V.-M. Kosma, H. Kujari, P. Luzi, G.M. Mariuzzi, E. Matze, R. Montironi, M. Scarpelli, D. Sierra, S. Sisti, S. Toikkanen, P. Tosi, W.F. Whimster, and E. Wisse. Standardized Mitotic Counts in Breast Cancer Evaluation of the Method. *Pathology - Research and Practice*, 192(9):931–941, January 1996.

[59] Tony J Collins. Imagej for microscopy. *Biotechniques*, 43(S1):S25–S30, 2007.

[60] Alberto Corvo, Marc A van Driel, and Michel A Westenberg. Pathova: A visual analytics tool for pathology diagnosis and reporting. In *2017 IEEE Workshop on Visual Analytics in Healthcare (VAHC)*, pages 77–83. IEEE, 2017.

[61] Ian A. Cree, Puay Hoon Tan, William D. Travis, Pieter Wesseling, Yukako Yagi, Valerie A. White, Dilani Lokuhetty, and Richard A. Scolyer. Counting mitoses: SI(ze) matters! *Modern Pathology*, 34(9):1651–1657, September 2021.

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

[63] Florian Daniel, Pavel Kucherbaev, Cinzia Cappiello, Boualem Benatallah, and Mohammad Allahbakhsh. Quality control in crowdsourcing: A survey of quality attributes, assessment techniques, and assurance actions. *ACM Comput. Surv.*, 51(1), jan 2018.

[64] Javier Del Ser, Alejandro Barredo-Arrieta, Natalia Díaz-Rodríguez, Francisco Herrera, Anna Saranti, and Andreas Holzinger. On generating trustworthy counterfactual explanations. *Information Sciences: an International Journal*, 655(C), February 2024.

[65] Claire Delong and Charles V Preuss. Black box warning. 2019.

[66] David Dov, Serge Assaad, Ameer Syedibrahim, Jonathan Bell, Jiaoti Huang, John Madden, Rex Bentley, Shannon McCall, Ricardo Henao, Lawrence Carin, and Wen-Chi Foo. 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.

[67] Eleonora Duregon, Adele Cassenti, Alessandra Pittaro, Laura Ventura, Rebecca Senetta, Roberta Rudà, and Paola Cassoni. Better see to better agree: phosphohistone H3 increases interobserver agreement in mitotic count for meningioma grading and imposes new specific thresholds. *Neuro-Oncology*, 17(5):663–669, 02 2015.

[68] Emir Efendić, Philippe PFM Van de Calseyde, and Anthony M Evans. Slow response times undermine trust in algorithmic (but not human) predictions. *Organizational Behavior and Human Decision Processes*, 157:103–114, 2020.

[69] Mehmet Günhan Ertosun and Daniel L Rubin. 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*, volume 2015, page 1899. American Medical Informatics Association, 2015.

[70] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, volume 96, pages 226–231, 1996.

[71] Andrew J Evans, Richard W Brown, Marilyn M Bui, Elizabeth A Chlipala, Christina Lacchetti, Danny A Milner Jr, Liron Pantanowitz, Anil V Parwani, Kearin Reid, Michael W Riben, et al. Validating whole slide imaging systems for diagnostic purposes in pathology: guideline update from the college of american pathologists in collaboration with the american society for clinical pathology and the association for pathology informatics. *Archives of pathology & laboratory medicine*, 146(4):440–450, 2022.

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

[73] Office of the FDA. Fda allows marketing of first whole slide imaging system for digital pathology.

[74] Office of the FDA. Fda authorizes software that can help identify prostate cancer.

[75] Paul M Fitts. The information capacity of the human motor system in controlling the amplitude of movement. *Journal of experimental psychology*, 47(6):381, 1954.

[76] Riccardo Fogliato, Shreya Chappidi, Matthew Lungren, Paul Fisher, Diane Wilson, Michael Fitzke, Mark Parkinson, Eric Horvitz, Kori Inkpen, and Besmira Nushi. Who goes first? influences of human-ai workflow on decision making in clinical imaging. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 1362–1374, New York, NY, USA, 2022. Association for Computing Machinery.

[77] Shintaro Fukushima, Mizuhiko Terasaki, Kiyohiko Sakata, Naohisa Miyagi, Seiya Kato, Yasuo Sugita, and Minoru Shigemori. Sensitivity and usefulness of anti-phosphohistone-h3 antibody immunostaining for counting mitotic figures in meningioma cases. *Brain tumor pathology*, 26:51–57, 2009.

[78] George W. Furnas and Benjamin B. Bederson. Space-scale diagrams: Understanding multiscale interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '95, page 234–241, USA, 1995. ACM Press/Addison-Wesley Publishing Co.

[79] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016.

[80] Susanne Gaube, Harini Suresh, Martina Raue, Alexander Merritt, Seth J Berkowitz, Eva Lermer, Joseph F Coughlin, John V Guttag, Errol Colak, and Marzyeh Ghassemi. Do as ai say: susceptibility in deployment of clinical decision-aids. *NPJ digital medicine*, 4(1):31, 2021.

[81] Jonathan R Genzen. An overview of united states physician training, certification, and career pathways in clinical pathology (laboratory medicine). *EJIFCC*, 24(1):21, 2013.

[82] Parmida Ghahremani, Yanyun Li, Arie Kaufman, Rami Vanguri, Noah Greenwald, Michael Angelo, Travis J Hollmann, and Saad Nadeem. Deepliif: Deep learning-inferred multiplex immunofluorescence for ihc image quantification. *bioRxiv*, 2021.

[83] 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. An analysis of pathologists' viewing processes as they diagnose whole slide digital images. *Journal of Pathology Informatics*, 13:100104, 2022.

[84] Michael Glueck, Tovi Grossman, and Daniel Wigdor. A model of navigation for very large data views. In *Proceedings of Graphics Interface 2013*, GI '13, page 9–16, CAN, 2013. Canadian Information Processing Society.

[85] Roland Goldbrunner, Pantelis Stavrinou, Michael D Jenkinson, Felix Sahm, Christian Mawrin, Damien C Weber, Matthias Preusser, Giuseppe Minniti, Morten Lund-Johansen, Florence Lefranc, Emanuel Houdart, Kita Sallabanda, Emilie Le Rhun, David Nieuwenhuizen, Ghazaleh Tabatabai, Riccardo Soffietti, and Michael Weller. EANO guideline on the diagnosis and management of meningiomas. *Neuro-Oncology*, 23(11):1821–1834, 06 2021.

[86] Simon Graham, Quoc Dang Vu, Shan E Ahmed Raza, Ayesha Azam, Yee Wah Tsang, Jin Tae Kwak, and Nasir Rajpoot. Hover-net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images. *Medical Image Analysis*, page 101563, 2019.

[87] Sorin Grigorescu, Bogdan Trasnea, Tiberiu Cocias, and Gigel Macesanu. A survey of deep learning techniques for autonomous driving. *Journal of Field Robotics*, 37(3):362–386, 2020.

[88] Hongyan Gu, Mohammad Haeri, Shuo Ni, Christopher Kazu Williams, Neda Zarrin-Khameh, Shino Magaki, and Xiang 'Anthony' Chen. Detecting mitoses with a convolutional neural network for midog 2022 challenge. In Bin Sheng and Marc Aubreville, editors, *Mitosis Domain Generalization and Diabetic Retinopathy Analysis*, pages 211–216, Cham, 2023. Springer Nature Switzerland.

[89] Hongyan Gu, Jingbin Huang, Lauren Hung, and Xiang 'Anthony' Chen. Lessons learned from designing an ai-enabled diagnosis tool for pathologists. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW1), apr 2021.

[90] Hongyan Gu, Yuan Liang, Yifan Xu, Christopher Kazu Williams, Shino Magaki, Negar Khanlou, Harry Vinters, Zesheng Chen, Shuo Ni, Chunxu Yang, Wenzhong Yan, Xinhai Robert Zhang, Yang Li, Mohammad Haeri, and Xiang 'Anthony' Chen. Improving workflow integration with xpath: Design and evaluation of a human-ai diagnosis system in pathology. *ACM Trans. Comput.-Hum. Interact.*, 30(2), mar 2023.

[91] Hongyan Gu, Chunxu Yang, Issa Al-kharouf, Shino Magaki, Nelli Lakis, Christopher Kazu Williams, Sallam Mohammad Alrosan, Ellie Kate Onstott, Wenzhong Yan, Negar Khanlou, Inma Cobos, Xinhai Robert Zhang, Neda Zarrin-Khameh, Harry V. Vinters, Xiang Anthony Chen, and Mohammad Haeri. Enhancing mitosis quantification and detection in meningiomas with computational digital pathology. *Acta Neuropathologica Communications*, 12(1):7, January 2024.

[92] Hongyan Gu, Chunxu Yang, Mohammad Haeri, Jing Wang, Shirley Tang, Wenzhong Yan, Shujin He, Christopher Kazu Williams, Shino Magaki, and Xiang 'Anthony' Chen. Augmenting pathologists with navipath: Design and evaluation of a human-ai collaborative navigation system. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA, 2023. Association for Computing Machinery.

[93] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[94] Sean Gustafson, Patrick Baudisch, Carl Gutwin, and Pourang Irani. Wedge: Clutter-free visualization of off-screen locations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '08, page 787–796, New York, NY, USA, 2008. Association for Computing Machinery.

[95] David A Gutman, Mohammed Khalilia, Sanghoon Lee, Michael Nalisnik, Zach Mullen, Jonathan Beezley, Deepak R Chittajallu, David Manthey, and Lee AD Cooper. The digital slide archive: A software platform for management, integration, and analysis of histology for cancer research. *Cancer research*, 77(21):e75–e78, 2017.

[96] Chu Han, Jiatai Lin, Jinhai Mai, Yi Wang, Qingling Zhang, Bingchao Zhao, Xin Chen, Xipeng Pan, Zhenwei Shi, Xiaowei Xu, et al. Multi-layer pseudo-supervision for histopathology tissue semantic segmentation using patch-level classification labels. *arXiv preprint arXiv:2110.08048*, 2021.

[97] Matthew G Hanna, Orly Ardon, Victor E Reuter, Sahussapont Joseph Sirintrapun, Christine England, David S Klimstra, and Meera R Hameed. Integrating digital pathology into clinical practice. *Modern Pathology*, 35(2):152–164, 2022.

[98] Matthew G Hanna, Victor E Reuter, Orly Ardon, David Kim, Sahussapont Joseph Sirintrapun, Peter J Schüffler, Klaus J Busam, Jennifer L Sauter, Edi Brogi, Lee K Tan, et al. Validation of a digital pathology system including remote review during the covid-19 pandemic. *Modern Pathology*, 33(11):2115–2127, 2020.

[99] Matthew G Hanna, Victor E Reuter, Meera R Hameed, Lee K Tan, Sarah Chiang, Carlie Sigel, Travis Hollmann, Dilip Giri, Jennifer Samboy, Carlos Moradel, et al. Whole slide imaging equivalency and efficiency study: experience at a large academic center. *Modern Pathology*, 32(7):916–928, 2019.

[100] Sandra G Hart and Lowell E Staveland. Development of nasa-tlx (task load index): Results of empirical and theoretical research. In *Advances in psychology*, volume 52, pages 139–183. Elsevier, 1988.

[101] 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. Similar image search for histopathology: Smily. *NPJ digital medicine*, 2(1):1–9, 2019.

[102] Tad Hirsch, Kritzia Merced, Shrikanth Narayanan, Zac E. Imel, and David C. Atkins. Designing contestability: Interaction design, machine learning, and mental health. In *Proceedings of the 2017 Conference on Designing Interactive Systems*, DIS '17, page 95–99, New York, NY, USA, 2017. Association for Computing Machinery.

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

[104] Andreas Holzinger, Georg Langs, Helmut Denk, Kurt Zatloukal, and Heimo Müller. Causability and explainability of artificial intelligence in medicine. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9(4):e1312, 2019.

[105] Eric Horvitz. Principles of mixed-initiative user interfaces. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pages 159–166, 1999.

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

[107] Zhi Huang, Federico Bianchi, Mert Yuksekgonul, Thomas J Montine, and James Zou. A visual–language foundation model for pathology image analysis using medical twitter. *Nature medicine*, 29(9):2307–2316, 2023.

[108] Zhi Huang, Eric Yang, Jeanne Shen, Dita Gratzinger, Frederick Eyerer, Brooke Liang, Jeffrey Nirschl, David Bingham, Alex M Dussaq, Christian Kunder, et al. A pathologist–ai collaboration framework for enhancing diagnostic accuracies and efficiencies. *Nature Biomedical Engineering*, pages 1–16, 2024.

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

[110] Aperio ImageScope. Aperio imagescope - pathology slide viewing software, Date Accessed: 2025-01-31.

[111] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

[112] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 590–597, 2019.

[113] 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. 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*, CHI '21, New York, NY, USA, 2021. Association for Computing Machinery.

[114] Maia Jacobs, Melanie F. Pradier, Thomas H. McCoy, Roy H. Perlis, Finale Doshi-Velez, and Krzysztof Z. Gajos. How machine-learning recommendations influence clinician treatment selections: the example of antidepressant selection. *Translational Psychiatry*, 11(1):108, February 2021.

[115] Jared Jessup, Robert Krueger, Simon Warchol, John Hoffer, Jeremy Muhlich, Cecily C Ritch, Giorgio Gaglia, Shannon Coy, Yu-An Chen, Jia-Ren Lin, et al. Scope2screen: Focus+ context techniques for pathology tumor assessment in multivariate image data. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):259–269, 2021.

[116] M. I. Jordan and T. M. Mitchell. Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245):255–260, 2015.

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

[118] Susanne Jul and George W. Furnas. Critical zones in desert fog: Aids to multiscale navigation. In *Proceedings of the 11th Annual ACM Symposium on User Interface Software and Technology*, UIST '98, page 97–106, New York, NY, USA, 1998. Association for Computing Machinery.

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

[120] Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. Interpreting interpretability: Understanding data scientists' use of interpretability tools for machine learning. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–14, New York, NY, USA, 2020. Association for Computing Machinery.

[121] Saif Khairat, David Marc, William Crosby, and Ali Al Sanousi. Reasons for physicians not adopting clinical decision support systems: Critical analysis. *JMIR Med Inform*, 6(2):e24, Apr 2018.

[122] Martin Köbel, Steve E Kalloger, Patricia M Baker, Carol A Ewanowich, Jocelyne Arseneau, Viktor Zherebitskiy, Soran Abdulkarim, Samuel Leung, Máire A Duggan, Dan Fontaine, et al. Diagnosis of ovarian carcinoma cell type is highly reproducible: a transcanadian study. *The American journal of surgical pathology*, 34(7):984–993, 2010.

[123] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 5338–5348. PMLR, 13–18 Jul 2020.

[124] Vivian Lai, Han Liu, and Chenhao Tan. "why is 'chicago' deceptive?" towards building model-driven tutorials for humans. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–13, New York, NY, USA, 2020. Association for Computing Machinery.

[125] Vivian Lai and Chenhao Tan. On human predictions with explanations and predictions of machine learning models: A case study on deception detection. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 29–38, New York, NY, USA, 2019. Association for Computing Machinery.

[126] Anastasia Lebedeva, Jaroslaw Kornowicz, Olesja Lammert, and Jörg Papenkordt. The role of response time for algorithm aversion in fast and slow thinking tasks. In *Artificial Intelligence in HCI: 4th International Conference, AI-HCI 2023, Held as Part of the 25th HCI International Conference, HCII 2023, Copenhagen, Denmark, July 23–28, 2023, Proceedings, Part I*, page 131–149, Berlin, Heidelberg, 2023. Springer-Verlag.

[127] Min Hun Lee, Daniel P. Siewiorek, Asim Smailagic, Alexandre Bernardino, and Sergi Bermúdez i Badia. 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*, CHI '21, New York, NY, USA, 2021. Association for Computing Machinery.

[128] Benedikt Leichtmann, Christina Humer, Andreas Hinterreiter, Marc Streit, and Martina Mara. Effects of explainable artificial intelligence on trust and human behavior in a high-risk decision task. *Computers in Human Behavior*, 139:107539, 2023.

[129] Christian Leistner, Amir Saffari, and Horst Bischof. Miforests: Multiple-instance learning with randomized trees. In *European Conference on Computer Vision*, pages 29–42. Springer, 2010.

[130] Joseph Carl Robnett Licklider. Man-computer symbiosis. *IRE transactions on human factors in electronics*, (1):4–11, 1960.

[131] Martin Lindvall, Claes Lundström, and Jonas Löwgren. Rapid assisted visual search: Supporting digital pathologists with imperfect ai. In *26th International Conference on Intelligent User Interfaces*, IUI '21, page 504–513, New York, NY, USA, 2021. Association for Computing Machinery.

[132] 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. 1399 h&e-stained sentinel lymph node sections of breast cancer patients: the camelyon dataset. *GigaScience*, 7(6):giy065, 2018.

[133] Geert Litjens, Clara I Sánchez, Nadya Timofeeva, Meyke Hermsen, Iris Nagtegaal, Iringo Kovacs, Christina Hulsbergen-Van De Kaa, Peter Bult, Bram Van Ginneken,

and Jeroen Van Der Laak. Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. *Scientific reports*, 6(1):26286, 2016.

[134] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. In *2008 Eighth IEEE International Conference on Data Mining*, pages 413–422. IEEE, 2008.

[135] Duri Long and Brian Magerko. What is ai literacy? competencies and design considerations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–16, New York, NY, USA, 2020. Association for Computing Machinery.

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

[137] 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. The 2021 WHO Classification of Tumors of the Central Nervous System: a summary. *Neuro-Oncology*, 23(8):1231–1251, 06 2021.

[138] Ming Y Lu, Bowen Chen, Drew FK Williamson, Richard J Chen, Melissa Zhao, Aaron K Chow, Kenji Ikemura, Ahrong Kim, Dimitra Pouli, Ankush Patel, et al. A multimodal generative ai copilot for human pathology. *Nature*, 634(8033):466–473, 2024.

[139] 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. 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):122–137, 2017.

[140] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

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

[142] Cedric Marchessoux, A. Nave Dufour, K. Espig, S. Monaco, A. Palekar, and L. Pantanowitz. Comparison display resolution on user impact for digital pathology. *Diagnostic Pathology*, 1(8), 2016.

[143] Anne L Martel, Dan Hosseinzadeh, Caglar Senaras, Yu Zhou, Azadeh Yazdanpanah, Rushin Shojaii, Emily S Patterson, Anant Madabhushi, and Metin N Gurcan. An image analysis resource for cancer research: Piip—pathology image informatics platform for visualization, analysis, and management. *Cancer research*, 77(21):e83–e86, 2017.

[144] Clare McGenity, Emily L Clarke, Charlotte Jennings, Gillian Matthews, Caroline Cartlidge, Deborah D Stocken, and Darren Treanor. Artificial intelligence in digital pathology: a systematic review and meta-analysis of diagnostic test accuracy. *npj Digital Medicine*, 7(1):114, 2024.

[145] Sara S McMillan, Michelle King, and Mary P Tully. How to use the nominal group and delphi techniques. *International journal of clinical pharmacy*, 38:655–662, 2016.

[146] John S Meyer, Consuelo Alvarez, Clara Milikowski, Neal Olson, Irma Russo, Jose Russo, Andrew Glass, Barbara A Zehnbauer, Karen Lister, and Reza Parwaresch. Breast carcinoma malignancy grading by Bloom–Richardson system vs proliferation index: reproducibility of grade and advantages of proliferation index. *Modern Pathology*, 18(8):1067–1078, August 2005.

[147] R. A. Miller and F. E. Masarie. The demise of the 'Greek Oracle' model for medical diagnostic systems, 1990.

[148] Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heintz, and Dan Roth. Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Comput. Surv.*, 56(2), sep 2023.

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

[150] Diana Montezuma, Ana Monteiro, João Fraga, Liliana Ribeiro, Sofia Gonçalves, André Tavares, João Monteiro, and Isabel Macedo-Pinto. Digital pathology implementation in private practice: specific challenges and opportunities. *Diagnostics*, 12(2):529, 2022.

[151] Katelyn Morrison, Donghoon Shin, Kenneth Holstein, and Adam Perer. Evaluating the impact of human explanation strategies on human-ai visual decision-making. *Proc. ACM Hum.-Comput. Interact.*, 7(CSCW1), apr 2023.

[152] Michael Nalisnik, Mohamed Amgad, Sanghoon Lee, Sameer H Halani, Jose Enrique Velazquez Vega, Daniel J Brat, David A Gutman, and Lee AD Cooper. Interactive phenotyping of large-scale histology imaging data with histomicsml. *Scientific reports*, 7(1):14588, 2017.

[153] MacLean P Nasrallah, Junhan Zhao, Cheng Che Tsai, David Meredith, Eliana Marostica, Keith L Ligon, Jeffrey A Golden, and Kun-Hsing Yu. Machine learning for cryosection pathology predicts the 2021 who classification of glioma. *Med*, 4(8):526–540, 2023.

[154] Mahsan Nourani, Chiradeep Roy, Jeremy E Block, Donald R Honeycutt, Tahrima Rahman, Eric Ragan, and Vibhav Gogate. Anchoring bias affects mental model formation and user reliance in explainable ai systems. In *26th International Conference on Intelligent User Interfaces*, IUI '21, page 340–350, New York, NY, USA, 2021. Association for Computing Machinery.

[155] OpenSeadragon. Openseadragon – an open-source, web-based viewer for high-resolution zoomable images, implemented in pure javascript, for desktop and mobile., Date Accessed: 2025-01-31.

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

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

[158] Joon Sung Park, Rick Barber, Alex Kirlik, and Karrie Karahalios. A slow algorithm improves users' assessments of the algorithm's accuracy. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW), nov 2019.

[159] Samir Passi and Mihaela Vorvoreanu. Overreliance on ai literature review. *Microsoft Research*, 2022.

[160] Ankush Patel, Ulysses GJ Balis, Jerome Cheng, Zaibo Li, Giovanni Lujan, David S McClintock, Liron Pantanowitz, and Anil Parwani. Contemporary whole slide imaging devices and their applications within the modern pathology department: a selected hardware review. *Journal of Pathology Informatics*, 12(1):50, 2021.

[161] Markus Plass, Michaela Kargl, Tim-Rasmus Kiehl, Peter Regitnig, Christian Geißler, Theodore Evans, Norman Zerbe, Rita Carvalho, Andreas Holzinger, and Heimo Müller. Explainability and causability in digital pathology. *The Journal of Pathology: Clinical Research*, 9(4):251–260, 2023. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/cjp2.322.

[162] Milda Pocevičiūtė, Gabriel Eilertsen, and Claes Lundström. Survey of xai in digital pathology. In *Artificial intelligence and machine learning for digital pathology*, pages 56–88. Springer, 2020.

[163] Alexander Rakhlin, Alexey Shvets, Vladimir Iglovikov, and Alexandr A Kalinin. Deep convolutional neural networks for breast cancer histology image analysis. In *International Conference Image Analysis and Recognition*, pages 737–744. Springer, 2018.

[164] Rebecca Randell, Thilina Ambepitiya, Claudia Mello-Thoms, Roy A Ruddle, David Brettle, Rhys G Thomas, and Darren Treanor. Effect of display resolution on time to diagnosis with virtual pathology slides in a systematic search task. *Journal of digital imaging*, 28(1):68–76, 2015.

[165] Rebecca Randell, Gordon Hutchins, John Sandars, Thilina Ambepitiya, Darren Treanor, Rhys Thomas, and Roy Ruddle. Using a high-resolution wall-sized virtual microscope to teach undergraduate medical students. In *CHI '12 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '12, page 2435–2440, New York, NY, USA, 2012. Association for Computing Machinery.

[166] Charvi Rastogi, Yunfeng Zhang, Dennis Wei, Kush R. Varshney, Amit Dhurandhar, and Richard Tomsett. Deciding fast and slow: The role of cognitive biases in ai-assisted decision-making. *Proc. ACM Hum.-Comput. Interact.*, 6(CSCW1), apr 2022.

[167] Saima Rathore, Tamim Niazi, Muhammad Aksam Iftikhar, and Ahmad Chaddad. Glioma grading via analysis of digital pathology images using machine learning. *Cancers*, 12(3):578, 2020.

[168] Peter Regitnig, Heimo Müller, and Andreas Holzinger. Expectations of Artificial Intelligence for Pathology. In Andreas Holzinger, Randy Goebel, Michael Mengel, and Heimo Müller, editors, *Artificial Intelligence and Machine Learning for Digital Pathology: State-of-the-Art and Future Challenges*, pages 1–15. Springer International Publishing, Cham, 2020.

[169] Jorge S Reis-Filho and Jakob Nikolas Kather. Overcoming the challenges to implementation of artificial intelligence in pathology. *JNCI: Journal of the National Cancer Institute*, 115(6):608–612, 2023.

[170] Juan Antonio Retamero, Jose Aneiros-Fernandez, and Raimundo G Del Moral. Complete digital pathology for routine histopathology diagnosis in a multicenter hospital network. *Archives of pathology & laboratory medicine*, 144(2):221–228, 2020.

[171] Daniel C. Robbins, Edward Cutrell, Raman Sarin, and Eric Horvitz. Zonezoom: Map navigation for smartphones with recursive view segmentation. In *Proceedings of the Working Conference on Advanced Visual Interfaces*, AVI '04, page 231–234, New York, NY, USA, 2004. Association for Computing Machinery.

[172] Thomas Roetzer-Pejrimovsky, Anna-Christina Moser, Baran Atli, Clemens Christian Vogel, Petra A Mercea, Romana Prihoda, Ellen Gelpi, Christine Haberler, Romana Höftberger, Johannes A Hainfellner, et al. The digital brain tumour atlas, an open histopathology resource. *Scientific Data*, 9(1):55, 2022.

[173] Ludovic Roux, Daniel Racoceanu, Nicolas Loménie, Maria Kulikova, Humayun Irshad, Jacques Klossa, Frédérique Capron, Catherine Genestie, Gilles Le Naour, and Metin N

Gurcan. Mitosis detection in breast cancer histological images an icpr 2012 contest. *Journal of pathology informatics*, 4, 2013.

[174] Roy A. Ruddle, Rhys G. Thomas, Rebecca Randell, Philip Quirke, and Darren Treanor. The design and evaluation of interfaces for navigating gigapixel images in digital pathology. *ACM Trans. Comput.-Hum. Interact.*, 23(1), jan 2016.

[175] Joel Saltz, Ashish Sharma, Ganesh Iyer, Erich Bremer, Feiqiao Wang, Alina Jasniewski, Tammy DiPrima, Jonas S Almeida, Yi Gao, Tianhao Zhao, et al. A containerized software system for generation, management, and exploration of features from whole slide tissue images. *Cancer research*, 77(21):e79–e82, 2017.

[176] Benzion Samueli, Natalie Aizenberg, Ruthy Shaco-Levy, Aviva Katzav, Yarden Kezerle, Judit Krausz, Salam Mazareb, Hagit Niv-Drori, Hila Belhanes Peled, Edmond Sabo, et al. Complete digital pathology transition: A large multi-center experience. *Pathology-Research and Practice*, 253:155028, 2024.

[177] Manojit Sarkar and Marc H. Brown. Graphical fisheye views of graphs. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '92, page 83–91, New York, NY, USA, 1992. Association for Computing Machinery.

[178] Mike Schaekermann, Graeme Beaton, Elaheh Sanoubari, Andrew Lim, Kate Larson, and Edith Law. Ambiguity-aware ai assistants for medical data analysis. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–14, New York, NY, USA, 2020. Association for Computing Machinery.

[179] Max Schemmer, Patrick Hemmer, Maximilian Nitsche, Niklas Kühl, and Michael Vössing. A meta-analysis of the utility of explainable artificial intelligence in human-ai decision-making. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '22, page 617–626, New York, NY, USA, 2022. Association for Computing Machinery.

[180] Max Schemmer, Niklas Kuehl, Carina Benz, Andrea Bartos, and Gerhard Satzger. Appropriate reliance on ai advice: Conceptualization and the effect of explanations. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, IUI '23, page 410–422, New York, NY, USA, 2023. Association for Computing Machinery.

[181] Caroline A Schneider, Wayne S Rasband, and Kevin W Eliceiri. Nih image to imagej: 25 years of image analysis. *Nature methods*, 9(7):671, 2012.

[182] Mark Sendak, Madeleine Clare Elish, Michael Gao, Joseph Futoma, William Ratliff, Marshall Nichols, Armando Bedoya, Suresh Balu, and Cara O'Brien. "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*, FAT* '20, page 99–109, New York, NY, USA, 2020. Association for Computing Machinery.

[183] B. Shneiderman. The eyes have it: a task by data type taxonomy for information visualizations. In *Proceedings 1996 IEEE Symposium on Visual Languages*, pages 336–343, 1996.

[184] Patrice Y Simard, Saleema Amershi, David M Chickering, Alicia Edelman Pelton, Soroush Ghorashi, Christopher Meek, Gonzalo Ramos, Jina Suh, Johan Verwey, Mo Wang, et al. Machine teaching: A new paradigm for building machine learning systems. *arXiv preprint arXiv:1707.06742*, 2017.

[185] Andrew H Song, Guillaume Jaume, Drew FK Williamson, Ming Y Lu, Anurag Vaidya, Tiffany R Miller, and Faisal Mahmood. Artificial intelligence for digital and computational pathology. *Nature Reviews Bioengineering*, 1(12):930–949, 2023.

[186] Robert Spence. Rapid, serial and visual: a presentation technique with potential. *Information visualization*, 1(1):13–19, 2002.

[187] Karin Stacke, Gabriel Eilertsen, Jonas Unger, and Claes Lundström. Measuring domain shift for deep learning in histopathology. *IEEE Journal of Biomedical and Health Informatics*, 25(2):325–336, 2021.

[188] 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. Evaluation of the use of combined artificial intelligence and pathologist assessment to review and grade prostate biopsies. *JAMA Network Open*, 3(11):e2023267–e2023267, 2020.

[189] Gregor Stiglic, Primoz Kocbek, Nino Fijacko, Marinka Zitnik, Katrien Verbert, and Leona Cilar. Interpretability of machine learning-based prediction models in healthcare. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(5):e1379, 2020.

[190] Eliza Strickland. Ibm watson, heal thyself: How ibm overpromised and underdelivered on ai health care. *IEEE Spectrum*, 56(4):24–31, 2019.

[191] 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. Artificial intelligence for diagnosis and grading of prostate cancer in biopsies: a population-based, diagnostic study. *The Lancet Oncology*, 21(2):222–232, 2020.

[192] Harry Surden. Artificial intelligence and law: An overview. *Georgia State University Law Review*, 35:19–22, 2019.

[193] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pages 6105–6114. PMLR, 2019.

[194] David Tellez, Maschenka Balkenhol, Irene Otte-Höller, Rob van de Loo, Rob Vogels, Peter Bult, Carla Wauters, Willem Vreuls, Suzanne Mol, Nico Karssemeijer, Geert Litjens, Jeroen van der Laak, and Francesco Ciompi. Whole-slide mitosis detection in h&e breast histology using phh3 as a reference to train distilled stain-invariant convolutional networks. *IEEE Transactions on Medical Imaging*, 37(9):2126–2136, 2018.

[195] Hamid Reza Tizhoosh and Liron Pantanowitz. Artificial intelligence and digital pathology: challenges and opportunities. *Journal of pathology informatics*, 9, 2018.

[196] Maciej Tomczak and Ewa Tomczak. The need to report effect size estimates revisited. an overview of some recommended measures of effect size. *Trends in sport sciences*, 1(21):19–25, 2014.

[197] Darren Treanor, Naomi Jordan-Owers, John Hodrien, Jason Wood, Phil Quirke, and Roy A Ruddle. Virtual reality powerwall versus conventional microscope for viewing pathology slides: an experimental comparison. *Histopathology*, 55(3):294–300, 2009.

[198] Darren Treanor and Phil Quirke. The virtual slide and conventional microscope-a direct comparison of their diagnostic efficiency. In *Annual Meeting of the Pathological Society of Great Britain and Ireland*, 2007.

[199] Philipp Tschandl, Christoph Rinner, Zoe Apalla, Giuseppe Argenziano, Noel Codella, Allan Halpern, Monika Janda, Aimilios Lallas, Caterina Longo, Josep Malvehy, et al. Human–computer collaboration for skin cancer recognition. *Nature Medicine*, 26(8):1229–1234, 2020.

[200] Stijn A. Van Bergeijk, Nikolas Stathonikos, Natalie D. Ter Hoeve, Maxime W. Lafarge, Tri Q. Nguyen, Paul J. Van Diest, and Mitko Veta. Deep learning supported mitoses counting on whole slide images: A pilot study for validating breast cancer grading in the clinical workflow. *Journal of Pathology Informatics*, 14:100316, 2023.

[201] Jeroen Van der Laak, Geert Litjens, and Francesco Ciompi. Deep learning in histopathology: the path to the clinic. *Nature medicine*, 27(5):775–784, 2021.

[202] Helena Vasconcelos, Matthew Jörke, Madeleine Grunde-McLaughlin, Tobias Gerstenberg, Michael S. Bernstein, and Ranjay Krishna. Explanations can reduce overreliance on ai systems during decision-making. *Proc. ACM Hum.-Comput. Interact.*, 7(CSCW1), apr 2023.

[203] Michael Veale and Frederik Zuiderveen Borgesius. Demystifying the draft eu artificial intelligence act—analysing the good, the bad, and the unclear elements of the proposed approach. *Computer Law Review International*, 22(4):97–112, 2021.

[204] 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. Predicting breast tumor proliferation from whole-slide images: the tupac16 challenge. *Medical image analysis*, 54:111–121, 2019.

[205] Mitko Veta, Paul J. Van Diest, Mehdi Jiwa, Shaimaa Al-Janabi, and Josien P. W. Pluim. Mitosis Counting in Breast Cancer: Object-Level Interobserver Agreement and Comparison to an Automatic Method. *PLOS ONE*, 11(8):e0161286, August 2016.

[206] 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. Assessment of algorithms for mitosis detection in breast cancer histopathology images. *Medical image analysis*, 20(1):237–248, 2015.

[207] Eugene Vorontsov, Alican Bozkurt, Adam Casson, George Shaikovski, Michal Zelechowski, Kristen Severson, Eric Zimmermann, James Hall, Neil Tenenholtz, Nicolo Fusi, et al. A foundation model for clinical-grade computational pathology and rare cancers detection. *Nature medicine*, pages 1–12, 2024.

[208] Brian Patrick Walcott, Brian V Nahed, Priscilla K Brastianos, and Jay S Loeffler. Radiation treatment for who grade ii and iii meningiomas. *Frontiers in oncology*, 3:227, 2013.

[209] Dakuo Wang, Elizabeth Churchill, Pattie Maes, Xiangmin Fan, Ben Shneiderman, Yuanchun Shi, and Qianying Wang. 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*, CHI EA '20, page 1–6, New York, NY, USA, 2020. Association for Computing Machinery.

[210] Dakuo Wang, Liuping Wang, Zhan Zhang, Ding Wang, Haiyi Zhu, Yvonne Gao, Xiangmin Fan, and Feng Tian. "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*, CHI '21, New York, NY, USA, 2021. Association for Computing Machinery.

[211] Dayong Wang, Aditya Khosla, Rishab Gargeya, Humayun Irshad, and Andrew H Beck. Deep learning for identifying metastatic breast cancer. *arXiv preprint arXiv:1606.05718*, 2016.

[212] Shidan Wang, Donghan M Yang, Ruichen Rong, Xiaowei Zhan, Junya Fujimoto, Hongyu Liu, John Minna, Ignacio Ivan Wistuba, Yang Xie, and Guanghua Xiao. Artificial intelligence in lung cancer pathology image analysis. *Cancers*, 11(11):1673, 2019.

[213] Weiwei Wang, Yuanshen Zhao, Lianghong Teng, Jing Yan, Yang Guo, Yuning Qiu, Yuchen Ji, Bin Yu, Dongling Pei, Wenchao Duan, et al. Neuropathologist-level integrated classification of adult-type diffuse gliomas using deep learning from whole-slide pathological images. *Nature Communications*, 14(1):6359, 2023.

[214] Xiyue Wang, Junhan Zhao, Eliana Marostica, Wei Yuan, Jietian Jin, Jiayu Zhang, Ruijiang Li, Hongping Tang, Kanran Wang, Yu Li, et al. A pathology foundation model for cancer diagnosis and prognosis prediction. *Nature*, 634(8035):970–978, 2024.

[215] Yinhai Wang, Kate E Williamson, Paul J Kelly, Jacqueline A James, and Peter W Hamilton. Surfaceslide: a multitouch digital pathology platform. *PloS one*, 7(1):e30783, 2012.

[216] Jeremy M Wolfe, Todd S Horowitz, Michael J Van Wert, Naomi M Kenner, Skyler S Place, and Nour Kibbi. Low target prevalence is a stubborn source of errors in visual search tasks. *Journal of experimental psychology: General*, 136(4):623, 2007.

[217] Yao Xie, Melody Chen, David Kao, Ge Gao, and Xiang'Anthony' Chen. Chexplain: Enabling physicians to explore and understand data-driven, ai-enabled medical imaging analysis. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2020.

[218] Hanwen Xu, Naoto Usuyama, Jaspreet Bagga, Sheng Zhang, Rajesh Rao, Tristan Naumann, Cliff Wong, Zelalem Gero, Javier González, Yu Gu, et al. A whole-slide foundation model for digital pathology from real-world data. *Nature*, pages 1–8, 2024.

[219] Cheng Xue, Qi Dou, Xueying Shi, Hao Chen, and Pheng-Ann Heng. Robust learning at noisy labeled medical images: Applied to skin lesion classification. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pages 1280–1283. IEEE, 2019.

[220] Qian Yang, Aaron Steinfeld, Carolyn Rosé, and John Zimmerman. Re-examining whether, why, and how human-ai interaction is uniquely difficult to design. In *Proceedings of the 2020 chi conference on human factors in computing systems*, pages 1–13, 2020.

[221] Qian Yang, Aaron Steinfeld, and John Zimmerman. Unremarkable ai: Fitting intelligent decision support into critical, clinical decision-making processes. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–11, 2019.

[222] Qian Yang, John Zimmerman, Aaron Steinfeld, Lisa Carey, and James F Antaki. 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*, pages 4477–4488, 2016.

[223] Jiawen Yao, Xinliang Zhu, Jitendra Jonnagaddala, Nicholas Hawkins, and Junzhou Huang. Whole slide images based cancer survival prediction using attention guided deep multiple instance learning networks. *Medical Image Analysis*, 65:101789, 2020.

[224] Polle T. Zellweger, Jock D. Mackinlay, Lance Good, Mark Stefik, and Patrick Baudisch. City lights: Contextual views in minimal space. In *CHI '03 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '03, page 838–839, New York, NY, USA, 2003. Association for Computing Machinery.

[225] David Y Zhang, Arsha Venkat, Hamdi Khasawneh, Rasoul Sali, Valerio Zhang, and Zhiheng Pei. Implementation of digital pathology and artificial intelligence in routine pathology practice. *Laboratory Investigation*, page 102111, 2024.

[226] Yongfeng Zhang, Guokun Lai, Min Zhang, Yi Zhang, Yiqun Liu, and Shaoping Ma. Explicit factor models for explainable recommendation based on phrase-level sentiment analysis. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, pages 83–92. ACM, 2014.

[227] Yunfeng Zhang, Q. Vera Liao, and Rachel K. E. Bellamy. Effect of confidence and explanation on accuracy and trust calibration in ai-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 295–305, New York, NY, USA, 2020. Association for Computing Machinery.

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

[229] Siqiong Zhou, Nicholaus Pfeiffer, Upala J. Islam, Imon Banerjee, Bhavika K. Patel, and Ashif S. Iquebal. Generating counterfactual explanations for causal inference in breast cancer treatment response. In *2022 IEEE 18th International Conference on Automation Science and Engineering (CASE)*, pages 955–960, 2022.

[230] Zhi-Hua Zhou. Multi-instance learning: A survey. *Department of Computer Science & Technology, Nanjing University, Tech. Rep*, 2004.