Augmenting Pathologists with NaviPath: Design and Evaluation of a Human-AI Collaborative Navigation System

Hongyan Gu
ghy@ucla.edu
University of California, Los Angeles
Los Angeles, California, USA

Chunxu Yang
chunxuyang@ucla.edu
University of California, Los Angeles
Los Angeles, California, USA

Mohammad Haeri
mhaeri@kumc.edu
University of Kansas Medical Center
Kansas City, Kansas, USA

Jing Wang
jingwang829@outlook.com
Beijing Tongren Hospital, Capital Medical University
Beijing, China

Shirley Tang
sjwtang@ucla.edu
University of California, Los Angeles
Los Angeles, California, USA

Wenzhong Yan
wzyan24@ucla.edu
University of California, Los Angeles
Los Angeles, California, USA

Shujin He
heshujun@mail.jnmc.edu.cn
Beijing Tongren Hospital, Capital Medical University
Beijing, China

Christopher Kazu Williams
ckwilliams@mednet.ucla.edu
UCLA David Geffen School of Medicine
Los Angeles, California, USA

Shino Magaki
smagaki@mednet.ucla.edu
UCLA David Geffen School of Medicine
Los Angeles, California, USA

Xiang 'Anthony' Chen
xac@ucla.edu
University of California, Los Angeles
Los Angeles, California, USA

ABSTRACT

Artificial Intelligence (AI) brings advancements to support pathologists in navigating high-resolution tumor images to search for pathology patterns of interest. However, existing AI-assisted tools have not realized this promised potential due to a lack of insight into pathology and HCI considerations for pathologists' navigation workflows in practice. We first conducted a formative study with six medical professionals in pathology to capture their navigation strategies. By incorporating our observations along with the pathologists' domain knowledge, we designed NaviPath — a human-AI collaborative navigation system. An evaluation study with 15 medical professionals in pathology indicated that: (i) compared to the manual navigation, participants saw more than twice the number of pathological patterns in unit time with NaviPath, and (ii) participants achieved higher precision and recall against the AI and the manual navigation on average. Further qualitative analysis revealed that navigation was more consistent with NaviPath, which can improve the overall examination quality.

CHI '23, April 23–28, 2023, Hamburg, Germany

ACM ISBN 978-1-4503-9421-5/23/04.
https://doi.org/10.1145/3544548.3580694

CCS CONCEPTS

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

KEYWORDS

Human-AI collaboration, digital pathology, navigation, medical AI

ACM Reference Format:
Hongyan Gu, Chunxu Yang, Mohammad Haeri, Jing Wang, Shirley Tang, Wenzhong Yan, Shujin He, Christopher Kazu Williams, Shino Magaki, and Xiang 'Anthony' Chen. 2023. 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), April 23–28, 2023, Hamburg, Germany.* ACM, New York, NY, USA, 19 pages. https://doi.org/10.1145/3544548.3580694

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 [55, 59], tumor specimens can be scanned into high-resolution digital scans, allowing medical professionals to access, analyze, and share these scans with digital interfaces [35, 53, 64]. However, literature has suggested that it might take longer for pathologists to examine digital scans compared to when using microscopes [36, 77]. 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 [63].

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) [6, 20, 34, 65, 86]. 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 [49]) have a low prevalence rate (<100/scan) and have extremely small dimensions compared to pathology scans (ratio up to 1:2000) [3]; (ii) pathologists require specific domain knowledge and navigation strategies [54, 63] 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 [33, 85]. 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 [44, 66, 84], radiology [16, 17] and pathology [12, 33, 47]. Motivated by the success of these advancements, this work 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 [23]. We summarized three observations that cross-validate the findings in previous research [30, 54, 63, 68]:

(1) **Overview first, then detail**: Pathologists followed this pattern of interacting with visual data as found in earlier works [30, 68]: 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 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 1(b)) to locate ROIs in low magnifications;
(3) **Low throughput in high magnifications**: Pathologists adopted a cautious and comprehensive navigation strategy (see Figure 1(c)) [54] 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 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 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 [86] and places short-cut navigation cues on the edge of the viewport (Figure 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.

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

[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) [3].

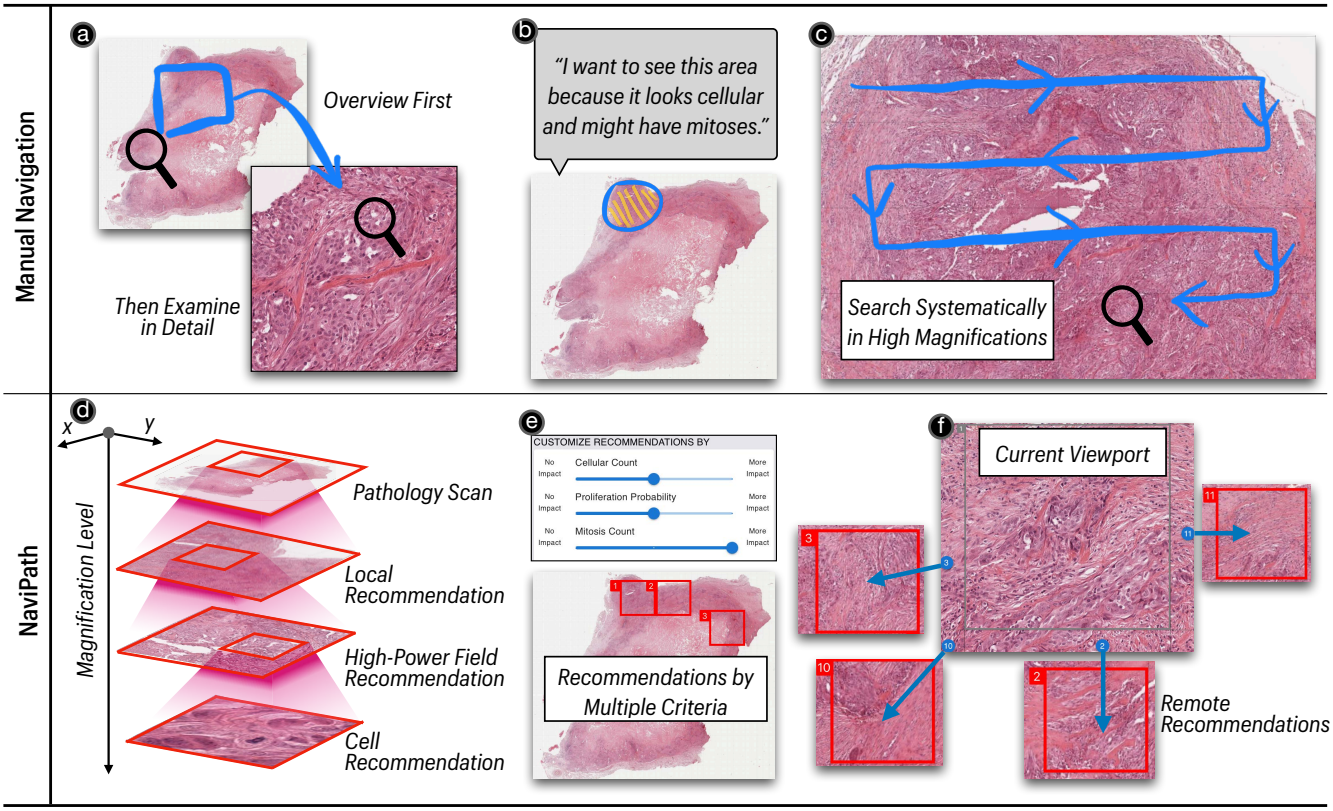


Figure 1: Comparison between pathologists' manual navigation in practice vs. NaviPath's designs. Observations on pathologists' manual navigation: (a) Pathologists usually overview a pathology scan with low magnifications, followed by switching to higher magnifications to examine regions of interest in 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.

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 work 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.

2 RELATED WORK

This section introduces three domains of work related to NaviPath: (i) interface designs to support pathologists' navigation, (ii) AI technologies for pathology, and (iii) human-AI collaboration to support medical decision-making.

2.1 Supporting Pathologists' Navigation with Interface Designs

Because the resolution of commercial off-the-shelf displays is significantly lower than pathology scans (up to 10^{12} pixels), intensive navigation is usually required for pathologists to search for features and make diagnoses [63]. Since the issue roots in resolution differences, one intuitive solution is to introduce displays with larger physical sizes and resolutions to pathologists [28, 61, 76, 82]. Literature has validated this solution, suggesting that pathologists utilized less pan and zoom interactions when using higher-resolution displays [52]. However, improving hardware requires purchasing costly, bulky, and specialized devices. And we believe that interface designs that can aid pathologists to work with high resolution digital scans are more closely related to what NaviPath achieves.

A recent study suggests that employing appropriate interface designs can accelerate pathologists' examination processes comparable to those of using the optical microscope [19]. Studies have well-explored designs to support navigating high-resolution images with limited size screens or displays [6, 34, 62, 65, 86]. Cockburn et al. summarized these into four categories: **focus + context (F+C)**, **overview + detail (O+D)**, zooming, and cue-based [20].

In digital pathology, the main-stream open-source [4, 21, 57, 67] and commercial [39] interfaces combine zooming and O+D designs, which include a zoomable canvas showing pathology scan details and an overview window that displays the thumbnail. Users can navigate high-resolution images with "pan and zoom" [29] interactions. However, criticisms suggest that such design demands a high mental effort and might be time-consuming [41, 63]. To compensate for the 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 [63]. Apart from O+D designs, Jessup et al. proposed an F+C interface for pathology image exploration [41]: a focal lens that magnifies the screen center and supports users' close-up examinations and explorations of multi-channeled pathology scans.

However, we argue that solely enhancing interface designs does not realize the full potential of digital scans. Because existing interface designs (without AI) lack support in assisting pathologists' visual searches [58], their navigation workflow can be substantially challenging while searching for small-sized, low-prevalence pathological patterns. Building upon the traditional O+D interface, the system proposed by this work adds AI and cue-based navigation, allowing pathologists to efficiently review AI findings with navigation cues.

2.2 AI Technologies for Pathology

Pathology has become an "attractive target" for applying AI because there exists a high variance in human diagnoses (i.e., the problem of consistency) and a shortage of trained pathologists (i.e., the issue of speed or efficiency) [71]. Driven by high demand, the past decade has experienced a burst of publicly annotated datasets that cover a broad range of pathology practices, from conducting high-level diagnostic tasks (e.g., identifying breast cancer metastasis [48], classifying kidney transplant biopsies [43]) to seeing low-level histopathological patterns (e.g., mitoses [3, 8, 50, 78]). Following the enrichment of datasets, numerous works have proposed deep learning models to perform pathology image analysis, with some achieving in-lab performance comparable to human pathologists [18, 24]. Furthermore, multiple works have applied deep learning models for mitosis detection, which include **Convolution Neural Networks (CNNs)** [31, 74], detection models (e.g., RetinaNet [3]), or a combination of both [45, 51]. However, unlike humans, research has indicated that current deep learning models have a generalizability limitation — their performance would deteriorate on the images with a domain shift (e.g., a shift caused by a difference in the data handling procedure in medical centers) [2, 70].

Setting aside the generalizability issue, the HCI problem of pathology using AI is its poor workflow integration: pathology is highly specialized domain in medicine, requiring specific expert knowledge and navigation strategies [54, 63] to facilitate doctors' examination. As state-of-the-art AI focuses on pushing the performance with data-driven, 'end-to-end' models, pathologists' needs for an AI's workflow integration is more or less ignored, which disincentives them from accepting and using AI in practice [85]. In this work, instead of employing AI to replace pathologists, we adapt AI closely to doctors' domain knowledge of navigation, enabling them to work collaboratively with AI. Our validation study shows that our human + AI approach is recognized to have a better workflow integration and can help pathologists achieve higher precision and recall on average compared to start-of-the-art AI.

2.3 Human-AI Collaboration for Medical Decision-Making

Similar to how humans work with others, the human-AI collaboration envisions humans and machines working symbiotically [46] to achieve mutual goals [79]. With the recent advancement of deep learning techniques, previous literature has established foundations of human-AI collaboration in the general domain (e.g., design [40] and content creation [25]). Furthermore, a number of HCI works have studied principles [38], guidelines [1], design recommendations [32], and information needs [13] to facilitate humans to work collaboratively with AI.

Following these pioneering works, research has investigated the broader applicability of 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 [7]. Wang et al. concluded the challenges of applying a clinical diagnostic support system in rural clinics [81]. 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 [44]. 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 [27]. Schaekermann et al. discovered that implementing ambiguity-aware AI was more effective in guiding medical experts' attention to contentious portions while reviewing sheep EEG data, compared to conventional AI [66]. Calisto et al. extended the designs of multi-modality radiology image viewing tools [14, 15]. 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 [16, 17].

Narrowing down to the pathology domain, promising works have employed a human + AI approach to support pathologists' examinations, bringing improvements in human errors [56, 80], between-subject agreements [11], time consumption [47], and mental workload [33]. For example, Lindvall et al. adapted the notion of **Rapid Serial Visual Presentation (RSVP)** [69] and developed a rapid assisted visual search system, allowing pathologists to see and adjust the AI-generated ROIs by sensitivity [47]. Gu et al. identified pathologists' challenges in practice and proposed a human-AI collaborative diagnosis system to perform multi-criteria, scan-level analysis for meningioma grading [33]. Notably, Cai et al. built a

pathology **content-based image retrieval (CBIR)** system with an imperfect model — pathologists could adjust the retrieved ROIs according to pathologist-defined concepts (e.g., stroma) to cope with AI imperfections [12].

Extending the exciting success of human-AI collaborative systems in pathology, this work continues to explore user-centered, integrable designs to embed AI assistance into pathologists' navigation processes. Specifically, going beyond presenting AI results to inform pathologists [12, 33], this work focuses on supporting the process with AI using designs that enable pathologists and AI to work symbiotically to navigate and gather information for diagnoses. Compared to previous human-AI navigation systems in pathology [47], NaviPath incorporates the domain knowledge of pathologists' navigation, which can improve the workflow integration and better augment pathologists' routines of using AI as a companion.

3 FORMATIVE STUDY & 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 the supplementary material for the 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 [3], 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.

3.1 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 [30, 68] and pathology [63]. 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., $\sim$ ×400 magnification). However, scanning the entire slide systematically in ×400 [54] can be substantially time-consuming because the field of view under ×400 is small compared to the pathology scan: a field of view under ×400 has a size of $0.16mm^2$, while a typical ×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 (<×50) as an overview. After that, they selected a few ROIs to proceed: for each ROI, they switched to medium-magnification ($\sim$ ×200) to maximize their fields-of-view while preserving cellular details. If a suspected cell was found, they would dive into high-magnification (×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 ×400) but also referred to macroscopic patterns (visible even in <×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 [42]). From the study, we observed that some participants preferred to use a cautious and comprehensive navigation strategy [54] (see Figure 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.

3.2 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 [2, 70], 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.

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.

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.

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' overviewing processes in low magnification. As shown in Figure 2(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 2(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 2(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 2(a)→(b), (b)→(c), (c)→(d)). Users may ignore the recommendation if an undesired one appears.

4.1.2 Customizable Recommendations by Multiple Criteria. NaviPath embeds pathologists' domain knowledge and employs three deep learning models (Figure 3(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 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 ×400 from ×200 views, NaviPath uses an EfficientNet-b3 model [73] to predict the proliferation probability — a criterion that relates to whether an ROI is likely to include mitosis, based on AI's impressions from ×200 magnification.
(3) **Mitosis Count**: Corresponding to pathologists' mitoses searching in ×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.

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 4(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 5(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 5(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 [32] and can help them focus on critical areas in limited time. Please refer to the supplementary material for the implementation details of AI models and the recommendation ranking algorithm.

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 [26], screen edges have infinite width, so it follows that.

Moreover, to increase pathologists' efficiency in seeing remote recommendations, NaviPath adapts the notion of citylight [86] by placing navigation cues on the edge of the interface (Figure 4(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.

[2]Specific sizes were justified by consulting with a board-certified pathologist (experience = 10 years)
[3]The size of one pixel is 0.25μm.

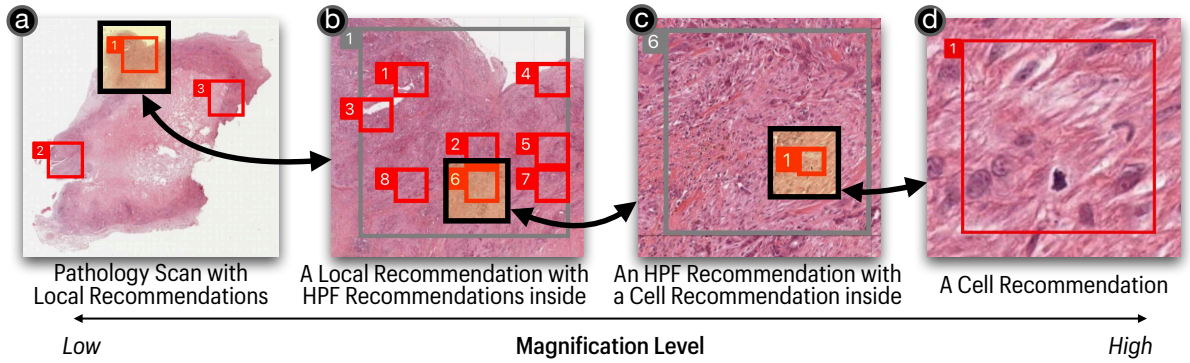


Figure 2: NaviPath generates hierarchical AI recommendations across multiple magnification levels: (a) Local recommendations (red boxes) lie in the lowest magnification, and can be seen directly on the pathology scan without zooming; (b) there are multiple High-Power Field (HPF) recommendations (red 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.

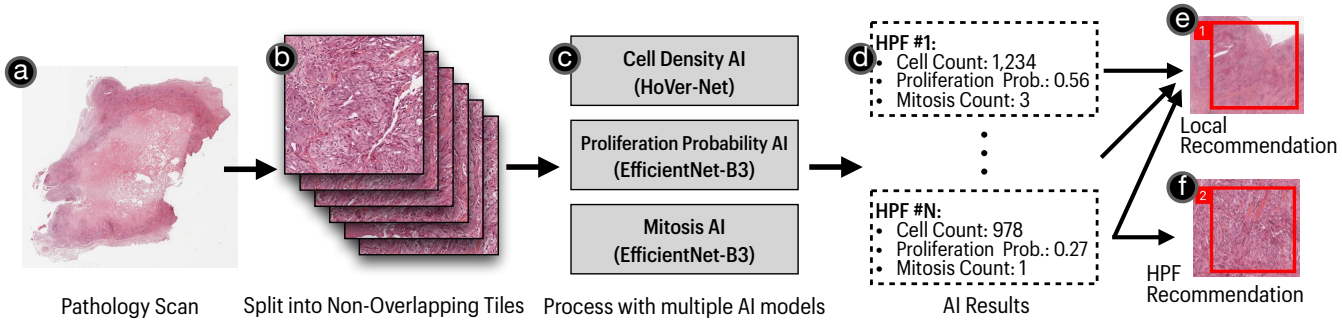


Figure 3: Generating Local and HPF recommendations with multiple criteria: (a) a pathology scan is first (b) split into non-overlapping tiles. Then, NaviPath uses (c) three AI models to analyze each tile to obtain (d) scores of cellular count, proliferation probability, and mitosis count. NaviPath will (e) aggregate scores from multiple tiles to generate Local recommendations, or (f) directly use these scores for HPF recommendations.

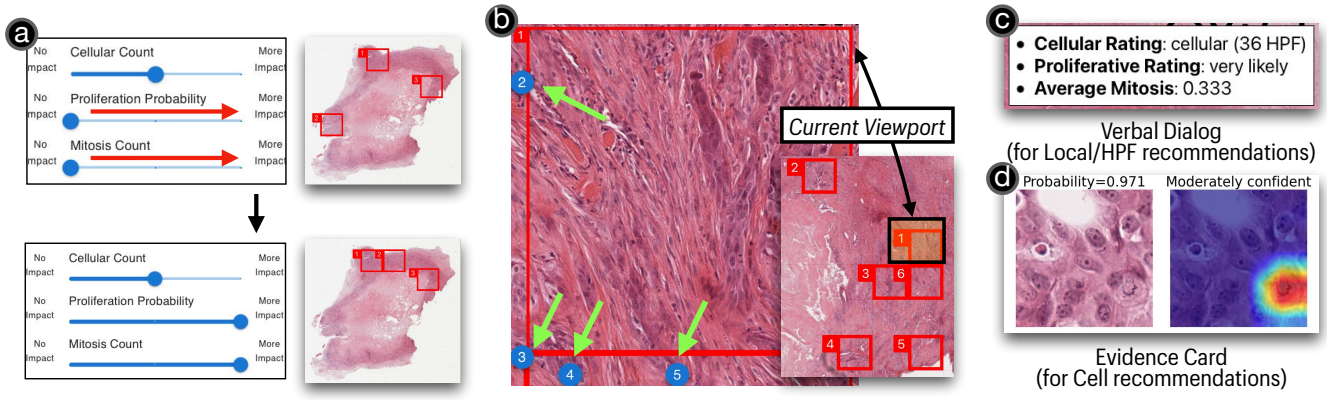


Figure 4: (a) NaviPath supports users to customize AI recommendations with a group of slide-bars: users can emphasize or rule out each of the three criteria (i.e., cellular count, proliferation probability, mitosis count) for NaviPath's recommendations; (b) NaviPath places navigation cues (pointed by 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.

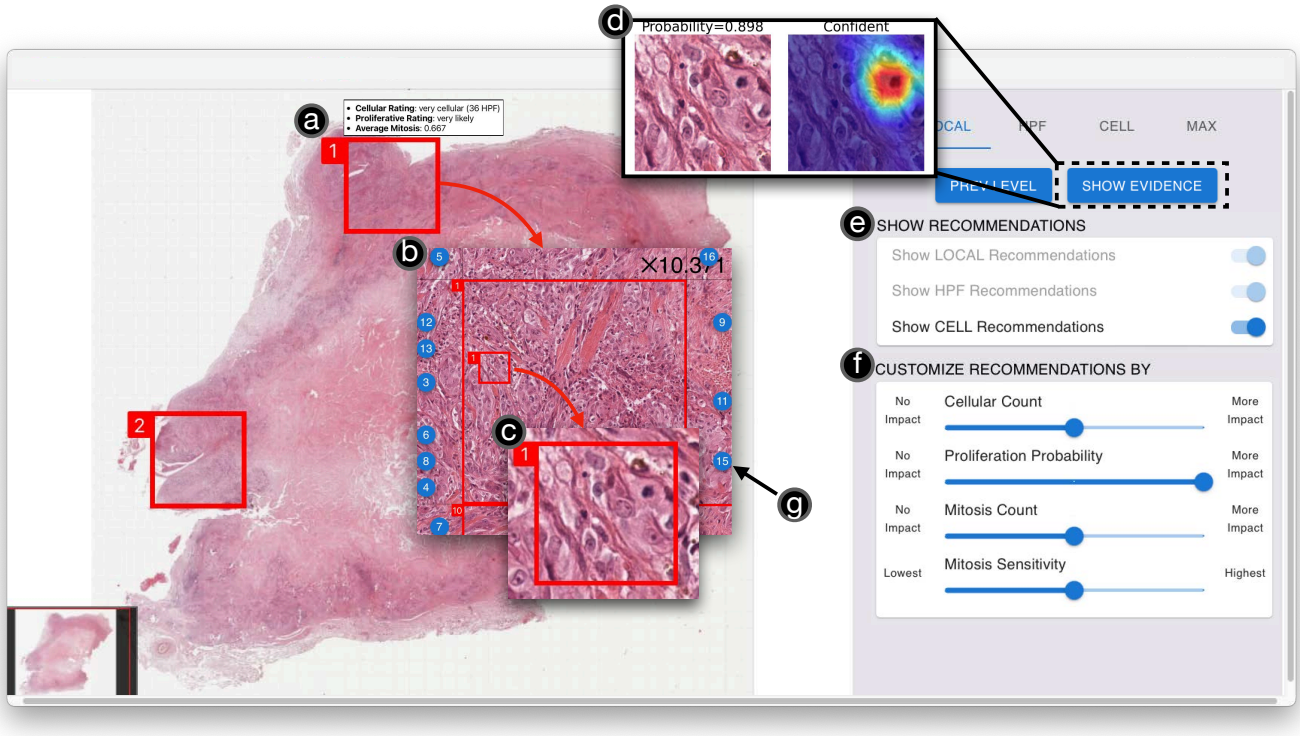


Figure 5: Overview of NaviPath's interface. (a) A Local recommendation (red box) with an explanation dialog. The number on the top-left corner represents the index of the recommendation (same for HPF and Cell recommendations); (b) An example of an HPF recommendation; (c) An example of a Cell recommendation; (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.

4.1.4 Explaining Each Recommendation. Since one criticism of deep learning models in pathology is that there is a lack of interpretability [72], **explainable AI (XAI)** techniques have been utilized to make AI "transparent, understandable and reliable" to pathologist users [60]. In NaviPath, we followed the suggestions from [32] 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 4(c)). The dialog helps users decide whether they should select and study recommended areas. Moreover, NaviPath adapts the design of a previous human-AI pathology system [33] and explains each Cell recommendation with an explanation card (Figure 4(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 supplementary material.

4.2 Navigating with NaviPath

A typical page of NaviPath is shown in Figure 5. A user's workflow in NaviPath starts by switching on (Figure 5e) and seeing Local recommendations (Figure 5a). 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 5b). In each HPF recommendation, users can continue to see Cell recommendations (Figure 5c) that show the precise locations of detected mitoses. For each Cell recommendation, users can view an explanation card on-demand (Figure 5d). After examining each HPF recommendation, users may click on the numbered navigation cue (Figure 5g) to hop to a remote HPF recommendation. During users' examination, they may customize the recommendations by interacting with a group of slide-bars (Figure 5f). Users' workflow ends when they are confident of signing out the case.

5 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 [3]. 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 [3]. For the cellular count calculation, we applied the model to 50 randomly-picked areas (size=512 × 512 pixels under ×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 [31], 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.

6 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;

6.1 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 1. 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 [39].

6.2 Data & Apparatus

We collected eight pathology scans of canine mammary carcinoma from a public dataset [3]. The average size of these scans was 7.15 giga-pixels. We acquired the ground-truth mitosis annotations from the same dataset [3]. Overall, the average **mitotic rate** (i.e., **MR**, mitotic count per unit area[4]) 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 1). 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.

6.3 Task & 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 1). 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[5], 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.

6.4 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 [63] and pathology AI [12], we investigated the research questions with the following measurements:

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

[4]https://www.cancer.gov/publications/dictionaries/cancer-terms/def/mitotic-rate
[5]https://documents.cap.org/protocols/cp-cns-18protocol-4000.pdf.

Table 1: Demographic information & 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			

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 2 Q1, Q2)) [37].

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 2 Q3), their confidence in the mitosis reportings (Table 2 Q4), attitudes toward using the system in the future (Table 2 Q5), and overall preference of system 1 vs. system 2 (Table 2 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 7) with a seven-scaled Likert question: (i) *"Is this feature useful to your examination?"* (1= Not useful at all → 7=Very useful); (ii) *"Compared to System 1, does this feature require extra effort?"*(1=No effort at all → 7=A lot of effort).

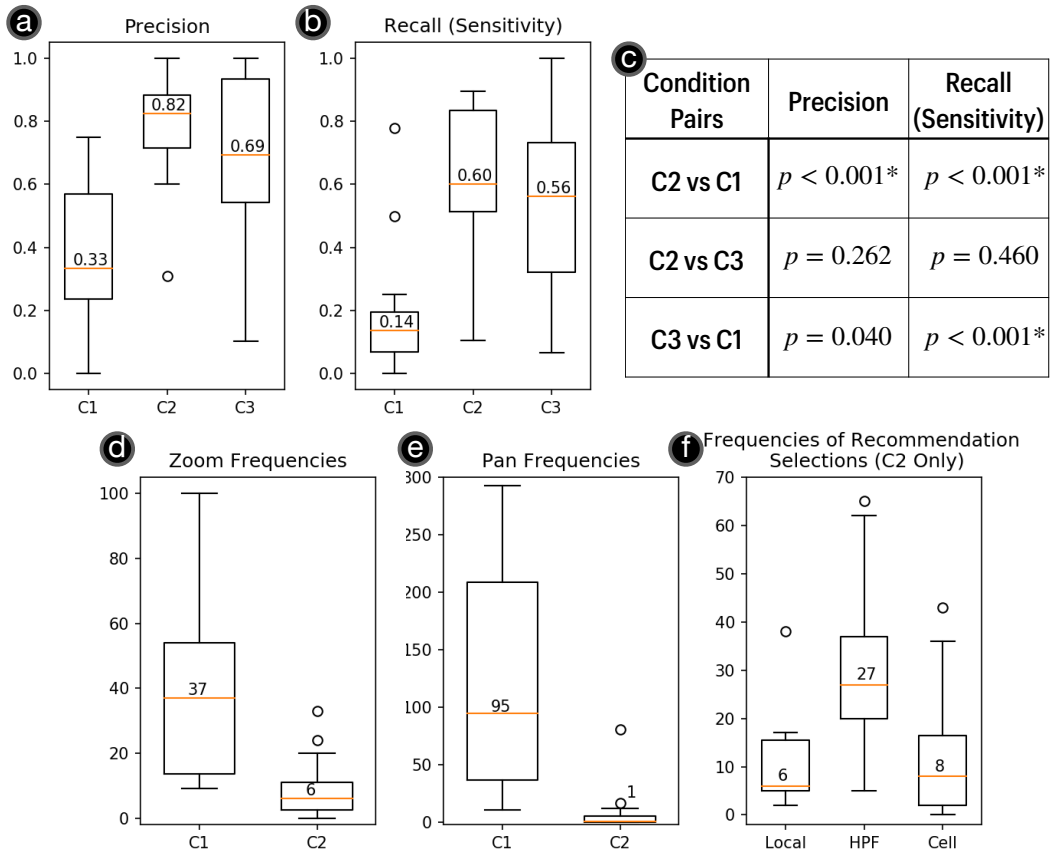
7 RESULT & FINDINGS

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.

7.1 Results for Research Questions

7.1.1 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 6(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[6]). 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 6(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

[6]The effect size of Kruskal-Wallis H-test η^2_H was calculated according to [75].



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 6: 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.

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 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 [83], 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 6(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.

Table 2: 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	

7.1.2 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 participants' examinations were significantly longer (p=0.09, effect size r=0.306[7], 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)

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 2 Q1). Furthermore, participants expressed less frustration using NaviPath (manual: μ =4.07, NaviPath: μ =2.40, p=0.024, r=0.412, Table 2 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.

7.1.3 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 6(d)). And the median pan interaction frequencies under **C1** and **C2** were 95 and 1 (Figure 6(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 2 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 2 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 2 Q5). Overall, as shown in Table 2 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

[7]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).

NaviPath was more capable of finding mitosis, had higher confidence while using NaviPath, and preferred to use it in the future.

7.2 Ratings on NaviPath's Components

To further understand whether each NaviPath component was useful for pathologists, we asked participants to rate each (see Figure 7). Here, we report the participants' ratings and discuss qualitative findings, organized by the categories of components:

7.2.1 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 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 7.1.3).

7.2.2 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 a previous study [33]. 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.

7.2.3 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 7.3.2.

7.2.4 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

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 7: 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).

components because participants spent extra effort comprehending the explanations.

7.3 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.

7.3.1 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 8(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 ground-truth mitosis density heatmap (Figure 8(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 8(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.

7.3.2 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 9(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 9(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 9(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 9(b), the bottom figure).
- **Cue-Based Hopping**: Participants clicked on the navigation cue to hop to a remote HPF recommendation (Figure 9c). 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 number 2."*(P10) As shown in Figure 9(c), the navigation trace is more irregular with cue-based hopping.

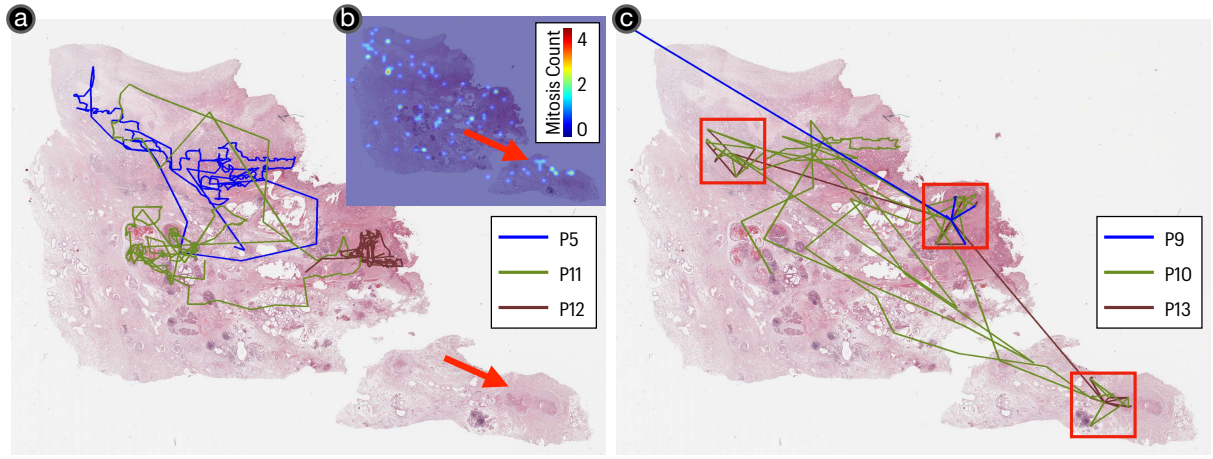


Figure 8: 2D projections of participants' traces with manual and NaviPath navigation on a pathology scan (zoom ignored). (a) Trace projections of P5, P11, and P12 with manual navigation. Note that all three participants did not examine the tissue on the bottom-right corner of the scan (pointed by the 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.

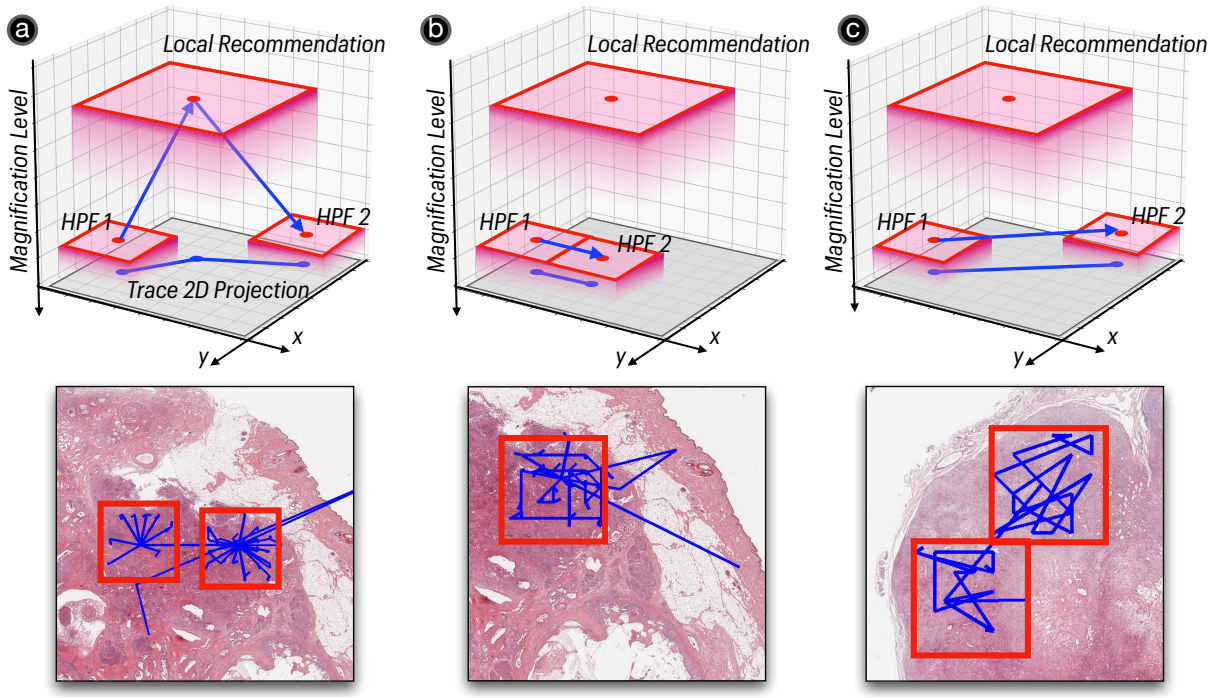


Figure 9: Three patterns of how our participants move to another HPF recommendation after examining one: (a) "Diving": first returned to the Local recommendation, overviewed the remaining HPF recommendations from the low magnification, and then dived down by selecting an HPF recommendation. The bottom 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.

8 DISCUSSION

8.1 Limitations

8.1.1 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.

8.1.2 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 [34] or Halo [5]) 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., [9]). 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 [22]. Finally, we also suggest future works to improve the predictability of medical AI, which we will discuss next.

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

8.2.1 Making AI-Enabled Systems Predictable. Previous work suggests that the disruptive behavior of AI might discourage medical professionals from using it in practice [85]. 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 [81]) 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 [12].

8.2.2 Balancing Simplicity and Informativeness. Doctors prefer simple, straightforward designs [32]. 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, Gu et al. indicated that pathology AI systems could provide levels of AI explanations for doctors: a simple, visual explanation was shown by default, while more detailed explanations could be retrieved on demand [33]. By balancing simplicity and informativeness, doctors can rapidly inquire about the most salient information with less confusion.

8.2.3 Decoupling Doctors and AI. Recent research has reported that utilizing AI may cause doctors' diagnoses to align with that of AI's [27]. 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 [7]. Moreover, previous research suggests that the domain gap in pathology image data will harm AI performance [2, 70]. 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 [10]. 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 [33]. Another unresolved question in this work 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 suggest readers see [33] for more detailed discussions.

9 CONCLUSION

This work 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. We hope insights provided by our solution can shed light on solving navigation challenges for other medical decision-making tasks.

ACKNOWLEDGMENTS

This work was funded in part by the Young Investigator Award by the Office of Naval Research and the National Science Foundation under grant IIS-1850183. We appreciate the anonymous participants for the user study, and the reviewers for their valuable feedback in improving the manuscript.

REFERENCES

[1] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. 2019. Guidelines for Human-AI Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3290605.3300233

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

[3] Marc Aubreville, Christof A. Bertram, Taryn A. Donovan, Christian Marzahl, Andreas Maier, and Robert Klopfleisch. 2020. A completely annotated whole slide image dataset of canine breast cancer to aid human breast cancer research. *Scientific Data* 7, 1 (Nov. 2020), 417. https://doi.org/10.1038/s41597-020-00756-z

[4] 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, Jacqueline A. James, Manuel Salto-Tellez, and Peter W. Hamilton. 2017. QuPath: Open source software for digital pathology image analysis. *Scientific Reports* 7, 1 (Dec. 2017), 16878. https://doi.org/10.1038/s41598-017-17204-5

[5] Patrick Baudisch and Ruth Rosenholtz. 2003. Halo: A Technique for Visualizing off-Screen Objects. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Ft. Lauderdale, Florida, USA) *(CHI '03)*. Association for Computing Machinery, New York, NY, USA, 481–488. https://doi.org/10.1145/642611.642695

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

[7] Emma Beede, Elizabeth Baylor, Fred Hersch, Anna Iurchenko, Lauren Wilcox, Paisan Ruamviboonsuk, and Laura M. Vardoulakis. 2020. A Human-Centered Evaluation of a Deep Learning System Deployed in Clinics for the Detection of Diabetic Retinopathy. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3313831.3376718

[8] Christof A. Bertram, Marc Aubreville, Christian Marzahl, Andreas Maier, and Robert Klopfleisch. 2019. A large-scale dataset for mitotic figure assessment on whole slide images of canine cutaneous mast cell tumor. *Scientific Data* 6, 1 (Nov. 2019), 274. https://doi.org/10.1038/s41597-019-0290-4

[9] Felix Bork, Christian Schnelzer, Ulrich Eck, and Nassir Navab. 2018. Towards Efficient Visual Guidance in Limited Field-of-View Head-Mounted Displays. *IEEE Transactions on Visualization and Computer Graphics* 24, 11 (2018), 2983–2992. https://doi.org/10.1109/TVCG.2018.2868584

[10] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. 2021. To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-Assisted Decision-Making. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1, Article 188 (apr 2021), 21 pages. https://doi.org/10.1145/3449287

[11] Wouter Bulten, Maschenka Balkenhol, Jean-Joël Awoumou Belinga, Américo Brilhante, Aslı Çakır, Lars Egevad, Martin Eklund, Xavier Farré, Katerina Geronatsiou, Vincent Molinié, Guilherme Pereira, Paromita Roy, Günter Saile, Paulo Salles, Ewout Schaafsma, Joëlle Tschui, Anne-Marie Vos, ISUP Pathology Imagebase Expert Panel, Hester van Boven, Robert Vink, Jeroen van der Laak, Christina Hulsbergen-van der Kaa, Geert Litjens, Brett Delahunt, Hemamali Samaratunga, David J. Grignon, Andrew J. Evans, Daniel M.Berney, Chin-Chen Pan, Glen Kristiansen, James G. Kench, Jon Oxley, Katia R.M. Leite, Jesse K. McKenney, Peter A. Humphrey, Samson W. Fine, Toyonori Tsuzuki, Murali Varma, Ming Zhou, Eva Comperat, David G. Bostwick, Kenneth A. Iczkowski, Cristina Magi-Galluzzi, John R. Srigley, Hiroyuki Takahashi, and Theo van der Kwast. 2021. Artificial intelligence assistance significantly improves Gleason grading of prostate biopsies by pathologists. *Modern Pathology* 34, 3 (2021), 660–671. https://doi.org/10.1038/s41379-020-0640-y

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

[13] Carrie J. Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. 2019. "Hello AI": Uncovering the Onboarding Needs of Medical Practitioners for Human-AI Collaborative Decision-Making. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 104 (nov 2019), 24 pages. https://doi.org/10.1145/3359206

[14] Francisco M. Calisto, Alfredo Ferreira, Jacinto C. Nascimento, and Daniel Gonçalves. 2017. Towards Touch-Based Medical Image Diagnosis Annotation. In *Proceedings of the 2017 ACM International Conference on Interactive Surfaces and Spaces* (Brighton, United Kingdom) *(ISS '17)*. Association for Computing Machinery, New York, NY, USA, 390–395. https://doi.org/10.1145/3132272.3134111

[15] Francisco Maria Calisto, Nuno Nunes, and Jacinto C. Nascimento. 2020. BreastScreening: On the Use of Multi-Modality in Medical Imaging Diagnosis. In *Proceedings of the International Conference on Advanced Visual Interfaces* (Salerno, Italy) *(AVI '20)*. Association for Computing Machinery, New York, NY, USA, Article 49, 5 pages. https://doi.org/10.1145/3399715.3399744

[16] Francisco Maria Calisto, Carlos Santiago, Nuno Nunes, and Jacinto C. Nascimento. 2021. Introduction of human-centric AI assistant to aid radiologists for multimodal breast image classification. *International Journal of Human-Computer Studies* 150 (2021), 102607. https://doi.org/10.1016/j.ijhcs.2021.102607

[17] Francisco Maria Calisto, Carlos Santiago, Nuno Nunes, and Jacinto C. Nascimento. 2022. BreastScreening-AI: Evaluating medical intelligent agents for human-AI interactions. *Artificial Intelligence in Medicine* 127 (2022), 102285. https://doi.org/10.1016/j.artmed.2022.102285

[18] Gabriele Campanella, Matthew G. Hanna, Luke Geneslaw, Allen Miraflor, Vitor Werneck Krauss Silva, Klaus J. Busam, Edi Brogi, Victor E. Reuter, David S. Klimstra, and Thomas J. Fuchs. 2019. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature Medicine* 25, 8 (Aug. 2019), 1301–1309. https://doi.org/10.1038/s41591-019-0508-1

[19] Emily Clarke, Daniel Doherty, Rebecca Randell, Jonathan Grek, Rhys Thomas, Roy A Ruddle, and Darren Treanor. 2022. Faster than light (microscopy): superiority of digital pathology over microscopy for assessment of immunohistochemistry. *Journal of Clinical Pathology* (Jan. 2022), jclinpath–2021–207961. https://doi.org/10.1136/jclinpath-2021-207961

[20] Andy Cockburn, Amy Karlson, and Benjamin B. Bederson. 2009. A Review of Overview+detail, Zooming, and Focus+context Interfaces. *ACM Comput. Surv.* 41, 1, Article 2 (jan 2009), 31 pages. https://doi.org/10.1145/1456650.1456652

[21] Tony J Collins. 2007. ImageJ for microscopy. *Biotechniques* 43, S1 (2007), S25–S30.

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

[23] Ian A. Cree, Puay Hoon Tan, William D. Travis, Pieter Wesseling, Yukako Yagi, Valerie A. White, Dilani Lokuhetty, and Richard A. Scolyer. 2021. Counting

mitoses: SI(ze) matters! *Modern Pathology* 34, 9 (Sept. 2021), 1651–1657. https://doi.org/10.1038/s41379-021-00825-7

[24] Babak Ehteshami Bejnordi, Mitko Veta, Paul Johannes van Diest, Bram van Ginneken, Nico Karssemeijer, Geert Litjens, Jeroen A. W. M. van der Laak, , and the CAMELYON16 Consortium. 2017. Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer. *JAMA* 318, 22 (12 2017), 2199–2210. https://doi.org/10.1001/jama.2017.14585

[25] Noyan Evirgen and Xiang 'Anthony' Chen. 2022. GANzilla: User-Driven Direction Discovery in Generative Adversarial Networks. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology* (Bend, OR, USA) *(UIST '22)*. Association for Computing Machinery, New York, NY, USA, Article 75, 10 pages. https://doi.org/10.1145/3526113.3545638

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

[27] Riccardo Fogliato, Shreya Chappidi, Matthew Lungren, Paul Fisher, Diane Wilson, Michael Fitzke, Mark Parkinson, Eric Horvitz, Kori Inkpen, and Besmira Nushi. 2022. Who Goes First? Influences of Human-AI Workflow on Decision Making in Clinical Imaging. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 1362–1374. https://doi.org/10.1145/3531146.3533193

[28] George W. Furnas and Benjamin B. Bederson. 1995. Space-Scale Diagrams: Understanding Multiscale Interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '95)*. ACM Press/Addison-Wesley Publishing Co., USA, 234–241. https://doi.org/10.1145/223904.223934

[29] George W. Furnas and Benjamin B. Bederson. 1995. Space-Scale Diagrams: Understanding Multiscale Interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '95)*. ACM Press/Addison-Wesley Publishing Co., USA, 234–241. https://doi.org/10.1145/223904.223934

[30] Michael Glueck, Tovi Grossman, and Daniel Wigdor. 2013. A Model of Navigation for Very Large Data Views. In *Proceedings of Graphics Interface 2013* (Regina, Sascatchewan, Canada) *(GI '13)*. Canadian Information Processing Society, CAN, 9–16.

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

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

[33] 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. 2022. Improving Workflow Integration with XPath: Design and Evaluation of a Human-AI Diagnosis System in Pathology. *ACM Trans. Comput.-Hum. Interact.* (dec 2022). https://doi.org/10.1145/3577011 Just Accepted.

[34] Sean Gustafson, Patrick Baudisch, Carl Gutwin, and Pourang Irani. 2008. Wedge: Clutter-Free Visualization of off-Screen Locations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Florence, Italy) *(CHI '08)*. Association for Computing Machinery, New York, NY, USA, 787–796. https://doi.org/10.1145/1357054.1357179

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

[36] Matthew G. Hanna, Victor E. Reuter, Meera R. Hameed, Lee K. Tan, Sarah Chiang, Carlie Sigel, Travis Hollmann, Dilip Giri, Jennifer Samboy, Carlos Moradel, Andrea Rosado, John R. Otilano, Christine England, Lorraine Corsale, Evangelos Stamelos, Yukako Yagi, Peter J. Schüffler, Thomas Fuchs, David S. Klimstra, and S. Joseph Sirintrapun. 2019. Whole slide imaging equivalency and efficiency study: experience at a large academic center. *Modern Pathology* 32, 7 (July 2019), 916–928. https://doi.org/10.1038/s41379-019-0205-0

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

[38] Eric Horvitz. 1999. Principles of Mixed-Initiative User Interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Pittsburgh, Pennsylvania, USA) *(CHI '99)*. Association for Computing Machinery, New York, NY, USA, 159–166. https://doi.org/10.1145/302979.303030

[39] Aperio ImageScope. Date Accessed: 2022-07-31. *Aperio ImageScope - Pathology Slide Viewing Software*. Aperio ImageScope. https://www.leicabiosystems.com/us/digital-pathology/manage/aperio-imagescope/

[40] Youngseung Jeon, Seungwan Jin, Patrick C. Shih, and Kyungsik Han. 2021. FashionQ: An AI-Driven Creativity Support Tool for Facilitating Ideation in Fashion Design. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 576, 18 pages. https://doi.org/10.1145/3411764.3445093

[41] Jared Jessup, Robert Krueger, Simon Warchol, John Hoffer, Jeremy Muhlich, Cecily C. Ritch, Giorgio Gaglia, Shannon Coy, Yu-An Chen, Jia-Ren Lin, Sandro Santagata, Peter K. Sorger, and Hanspeter Pfister. 2022. Scope2Screen: Focus+Context Techniques for Pathology Tumor Assessment in Multivariate Image Data. *IEEE Transactions on Visualization and Computer Graphics* 28, 1 (2022), 259–269. https://doi.org/10.1109/TVCG.2021.3114786

[42] Susanne Jul and George W. Furnas. 1998. Critical Zones in Desert Fog: Aids to Multiscale Navigation. In *Proceedings of the 11th Annual ACM Symposium on User Interface Software and Technology* (San Francisco, California, USA) *(UIST '98)*. Association for Computing Machinery, New York, NY, USA, 97–106. https://doi.org/10.1145/288392.288578

[43] Jesper Kers, Roman D Bülow, Barbara M Klinkhammer, Gerben E Breimer, Francesco Fontana, Adeyemi Adefidipe Abiola, Rianne Hofstraat, Garry L Corthals, Hessel Peters-Sengers, Sonja Djudjaj, Saskia von Stillfried, David L Hölscher, Tobias T Pieters, Arjan D van Zuilen, Frederike J Bemelman, Azam S Nurmohamed, Maarten Naesens, Joris J T H Roelofs, Sandrine Florquin, Jürgen Floege, Tri Q Nguyen, Jakob N Kather, and Peter Boor. 2022. Deep learning-based classification of kidney transplant pathology: a retrospective, multicentre, proof-of-concept study. *The Lancet Digital Health* 4, 1 (Jan. 2022), e18–e26. https://doi.org/10.1016/S2589-7500(21)00211-9

[44] Min Hun Lee, Daniel P. Siewiorek, Asim Smailagic, Alexandre Bernardino, and Sergi Bermúdez i Badia. 2021. A Human-AI Collaborative Approach for Clinical Decision Making on Rehabilitation Assessment. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 392, 14 pages. https://doi.org/10.1145/3411764.3445472

[45] Chao Li, Xinggang Wang, Wenyu Liu, and Longin Jan Latecki. 2018. DeepMitosis: Mitosis detection via deep detection, verification and segmentation networks. *Medical image analysis* 45 (2018), 121–133.

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

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

[48] Geert Litjens, Peter Bandi, Babak Ehteshami Bejnordi, Oscar Geessink, Maschenka Balkenhol, Peter Bult, Altuna Halilovic, Meyke Hermsen, Rob van de Loo, Rob Vogels, Quirine F Manson, Nikolas Stathonikos, Alexi Baidoshvili, Paul van Diest, Carla Wauters, Marcory van Dijk, and Jeroen van der Laak. 2018. 1399 H&E-stained sentinel lymph node sections of breast cancer patients: the CAMELYON dataset. *GigaScience* 7, 6 (05 2018). https://doi.org/10.1093/gigascience/giy065 arXiv:https://academic.oup.com/gigascience/article-pdf/7/6/giy065/25045140/giy065_reviewer_2_report_(revision_1).pdf giy065.

[49] David N Louis, Arie Perry, Pieter Wesseling, Daniel J Brat, Ian A Cree, Dominique Figarella-Branger, Cynthia Hawkins, HK Ng, Stefan M Pfister, Guido Reifenberger, et al. 2021. The 2021 WHO classification of tumors of the central nervous system: a summary. *Neuro-oncology* 23, 8 (2021), 1231–1251. https://doi.org/10.1093/neuonc/noab106

[50] Roux Ludovic, Racoceanu Daniel, Loménie Nicolas, Kulikova Maria, Irshad Humayun, Klossa Jacques, Capron Frédérique, Genestie Catherine, Le Naour Gilles, and Gurcan Metin N. 2013. Mitosis detection in breast cancer histological images An ICPR 2012 contest. *Journal of Pathology Informatics* 4, 1 (2013), 8. https://doi.org/10.4103/2153-3539.112693

[51] Tahir Mahmood, Muhammad Arsalan, Muhammad Owais, Min Beom Lee, and Kang Ryoung Park. 2020. Artificial intelligence-based mitosis detection in breast cancer histopathology images using faster R-CNN and deep CNNs. *Journal of clinical medicine* 9, 3 (2020), 749.

[52] Cedric Marchessoux, A. Nave Dufour, K. Espig, S. Monaco, A. Palekar, and L. Pantanowitz. 2016. Comparison Display Resolution On User Impact For Digital Pathology. *Diagnostic Pathology* 1, 8 (2016). https://doi.org/10.17629/www.diagnosticpathology.eu-2016-8:168

[53] Anne L Martel, Dan Hosseinzadeh, Caglar Senaras, Yu Zhou, Azadeh Yazdanpanah, Rushin Shojaii, Emily S Patterson, Anant Madabhushi, and Metin N Gurcan. 2017. An image analysis resource for cancer research: PIIP—pathology image informatics platform for visualization, analysis, and management. *Cancer research* 77, 21 (2017), e83–e86. https://doi.org/10.1158/0008-5472.CAN-17-0629

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

[55] Office of the FDA. Date Accessed: 2022-07-31. *FDA allows marketing of first whole slide imaging system for Digital Pathology*. https://www.fda.gov/news-

events/press-announcements/fda-allows-marketing-first-whole-slide-imaging-system-digital-pathology

[56] Office of the FDA. Date Accessed: 2022-07-31. *FDA Authorizes Software that Can Help Identify Prostate Cancer.* https://www.fda.gov/news-events/press-announcements/fda-authorizes-software-can-help-identify-prostate-cancer

[57] OpenSeadragon. Date Accessed: 2022-07-31. *OpenSeadragon – An open-source, web-based viewer for high-resolution zoomable images, implemented in pure JavaScript, for desktop and mobile.* https://openseadragon.github.io/

[58] John Palmer. 1995. Attention in visual search: Distinguishing four causes of a set-size effect. *Current directions in psychological science* 4, 4 (1995), 118–123.

[59] Liron Pantanowitz, Paul N. Valenstein, Andrew J. Evans, Keith J. Kaplan, John D. Pfeifer, David C. Wilbur, Laura C. Collins, and Terence J. Colgan. 2011. Review of the current state of whole slide imaging in pathology. *Journal of Pathology Informatics* 2, 1 (2011), 36. https://doi.org/10.4103/2153-3539.83746

[60] Milda Pocevičiūtė, Gabriel Eilertsen, and Claes Lundström. 2020. *Survey of XAI in Digital Pathology.* Springer International Publishing, Cham, 56–88. https://doi.org/10.1007/978-3-030-50402-1_4

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

[62] Daniel C. Robbins, Edward Cutrell, Raman Sarin, and Eric Horvitz. 2004. ZoneZoom: Map Navigation for Smartphones with Recursive View Segmentation. In *Proceedings of the Working Conference on Advanced Visual Interfaces* (Gallipoli, Italy) *(AVI '04)*. Association for Computing Machinery, New York, NY, USA, 231–234. https://doi.org/10.1145/989863.989901

[63] Roy A. Ruddle, Rhys G. Thomas, Rebecca Randell, Philip Quirke, and Darren Treanor. 2016. The Design and Evaluation of Interfaces for Navigating Gigapixel Images in Digital Pathology. *ACM Trans. Comput.-Hum. Interact.* 23, 1, Article 5 (jan 2016), 29 pages. https://doi.org/10.1145/2834117

[64] Joel Saltz, Ashish Sharma, Ganesh Iyer, Erich Bremer, Feiqiao Wang, Alina Jasniewski, Tammy DiPrima, Jonas S. Almeida, Yi Gao, Tianhao Zhao, Mary Saltz, and Tahsin Kurc. 2017. A Containerized Software System for Generation, Management, and Exploration of Features from Whole Slide Tissue Images. *Cancer Research* 77, 21 (Nov. 2017), e79–e82. https://doi.org/10.1158/0008-5472.CAN-17-0316

[65] Manojit Sarkar and Marc H. Brown. 1992. Graphical Fisheye Views of Graphs. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Monterey, California, USA) *(CHI '92)*. Association for Computing Machinery, New York, NY, USA, 83–91. https://doi.org/10.1145/142750.142763

[66] Mike Schaekermann, Graeme Beaton, Elaheh Sanoubari, Andrew Lim, Kate Larson, and Edith Law. 2020. Ambiguity-Aware AI Assistants for Medical Data Analysis. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376506

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

[68] B. Shneiderman. 1996. The eyes have it: a task by data type taxonomy for information visualizations. In *Proceedings 1996 IEEE Symposium on Visual Languages*. 336–343. https://doi.org/10.1109/VL.1996.545307

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

[70] Karin Stacke, Gabriel Eilertsen, Jonas Unger, and Claes Lundström. 2020. Measuring domain shift for deep learning in histopathology. *IEEE journal of biomedical and health informatics* 25, 2 (2020), 325–336.

[71] David F Steiner, Po-Hsuan Cameron Chen, and Craig H Mermel. 2021. Closing the translation gap: AI applications in digital pathology. *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer* 1875, 1 (2021), 188452.

[72] Gregor Stiglic, Primoz Kocbek, Nino Fijacko, Marinka Zitnik, Katrien Verbert, and Leona Cilar. 2020. Interpretability of machine learning-based prediction models in healthcare. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 10, 5 (2020), e1379.

[73] Mingxing Tan and Quoc Le. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*. PMLR, 6105–6114.

[74] 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. 2018. 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 (2018), 2126–2136. https://doi.org/10.1109/TMI.2018.2820199

[75] Maciej Tomczak and Ewa Tomczak. 2014. The need to report effect size estimates revisited. An overview of some recommended measures of effect size. *Trends in sport sciences* 1, 21 (2014), 19–25.

[76] Darren Treanor, Naomi Jordan-Owers, John Hodrien, Jason Wood, Phil Quirke, and Roy A Ruddle. 2009. Virtual reality Powerwall versus conventional microscope for viewing pathology slides: an experimental comparison. *Histopathology* 55, 3 (2009), 294–300.

[77] Darren Treanor and Phil Quirke. 2007. 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.*

[78] Mitko Veta, Yujing J. Heng, Nikolas Stathonikos, Babak Ehteshami Bejnordi, Francisco Beca, Thomas Wollmann, Karl Rohr, Manan A. Shah, Dayong Wang, Mikael Rousson, Martin Hedlund, David Tellez, Francesco Ciompi, Erwan Zerhouni, David Lanyi, Matheus Viana, Vassili Kovalev, Vitali Liauchuk, Hady Ahmady Phoulady, Talha Qaiser, Simon Graham, Nasir Rajpoot, Erik Sjöblom, Jesper Molin, Kyunghyun Paeng, Sangheum Hwang, Sunggyun Park, Zhipeng Jia, Eric I-Chao Chang, Yan Xu, Andrew H. Beck, Paul J. van Diest, and Josien P.W. Pluim. 2019. Predicting breast tumor proliferation from whole-slide images: The TUPAC16 challenge. *Medical Image Analysis* 54 (2019), 111–121. https://doi.org/10.1016/j.media.2019.02.012

[79] Dakuo Wang, Elizabeth Churchill, Pattie Maes, Xiangmin Fan, Ben Shneiderman, Yuanchun Shi, and Qianying Wang. 2020. From Human-Human Collaboration to Human-AI Collaboration: Designing AI Systems That Can Work Together with People. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI EA '20)*. Association for Computing Machinery, New York, NY, USA, 1–6. https://doi.org/10.1145/3334480.3381069

[80] Dayong Wang, Aditya Khosla, Rishab Gargeya, Humayun Irshad, and Andrew H Beck. 2016. Deep learning for identifying metastatic breast cancer. *arXiv preprint arXiv:1606.05718* (2016).

[81] Dakuo Wang, Liuping Wang, Zhan Zhang, Ding Wang, Haiyi Zhu, Yvonne Gao, Xiangmin Fan, and Feng Tian. 2021. "Brilliant AI Doctor" in Rural Clinics: Challenges in AI-Powered Clinical Decision Support System Deployment. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 697, 18 pages. https://doi.org/10.1145/3411764.3445432

[82] Yinhai Wang, Kate E Williamson, Paul J Kelly, Jacqueline A James, and Peter W Hamilton. 2012. SurfaceSlide: a multitouch digital pathology platform. *PloS one* 7, 1 (2012), e30783.

[83] Jeremy M Wolfe, Todd S Horowitz, Michael J Van Wert, Naomi M Kenner, Skyler S Place, and Nour Kibbi. 2007. Low target prevalence is a stubborn source of errors in visual search tasks. *Journal of experimental psychology: General* 136, 4 (2007), 623.

[84] Qian Yang, Aaron Steinfeld, and John Zimmerman. 2019. 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* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–11. https://doi.org/10.1145/3290605.3300468

[85] Qian Yang, John Zimmerman, Aaron Steinfeld, Lisa Carey, and James F. Antaki. 2016. Investigating the Heart Pump Implant Decision Process: Opportunities for Decision Support Tools to Help. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 4477–4488. https://doi.org/10.1145/2858036.2858373

[86] Polle T. Zellweger, Jock D. Mackinlay, Lance Good, Mark Stefik, and Patrick Baudisch. 2003. City Lights: Contextual Views in Minimal Space. In *CHI '03 Extended Abstracts on Human Factors in Computing Systems* (Ft. Lauderdale, Florida, USA) *(CHI EA '03)*. Association for Computing Machinery, New York, NY, USA, 838–839. https://doi.org/10.1145/765891.766022