Int. J. Human–Computer Studies 190 (2024) 103315

Contents lists available at ScienceDirect

International Journal of Human - Computer Studies

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

Majority voting of doctors improves appropriateness of AI reliance in pathology

Hongyan Gu [a,*], Chunxu Yang [a], Shino Magaki [b], Neda Zarrin-Khameh [c], Nelli S. Lakis [d], Inma Cobos [e], Negar Khanlou [b], Xinhai R. Zhang [f], Jasmeet Assi [d], Joshua T. Byers [g], Ameer Hamza [d], Karam Han [h], Anders Meyer [d], Hilda Mirbaha [b], Carrie A. Mohila [i], Todd M. Stevens [d], Sara L. Stone [j], Wenzhong Yan [a], Mohammad Haeri [d,*], Xiang 'Anthony' Chen [a]

[a] Department of Electrical and Computer Engineering, University of California, Los Angeles, Los Angeles, USA
[b] Department of Pathology and Laboratory Medicine, UCLA David Geffen School of Medicine, Los Angeles, USA
[c] Department of Pathology and Laboratory Medicine, Baylor College of Medicine and Ben Taub Hospital, Houston, USA
[d] Department of Pathology and Laboratory Medicine, University of Kansas Medical Center, Kansas City, USA
[e] Department of Pathology, Stanford School of Medicine, Stanford, USA
[f] McGovern Medical School, University of Texas Health Science Center at Houston, Houston, USA
[g] School of Medicine, University of California, San Francisco, San Francisco, USA
[h] Department of Pathology and Laboratory Medicine, University of Wisconsin-Madison, Madison, USA
[i] Department of Pathology, Baylor College of Medicine and Texas Children's Hospital, Houston, USA
[j] Department of Pathology, Hospital of the University of Pennsylvania, Philadelphia, USA

ARTICLE INFO

Keywords:
Appropriate reliance
Artificial intelligence
Majority voting
Pathology

ABSTRACT

As Artificial Intelligence (AI) making advancements in medical decision-making, there is a growing need to ensure doctors develop appropriate reliance on AI to avoid adverse outcomes. However, existing methods in enabling appropriate AI reliance might encounter challenges while being applied in the medical domain. With this regard, this work employs and provides the validation of an alternative approach – majority voting – to facilitate appropriate reliance on AI in medical decision-making. This is achieved by a multi-institutional user study involving 32 medical professionals with various backgrounds, focusing on the pathology task of visually detecting a pattern, mitoses, in tumor images. Here, the majority voting process was conducted by synthesizing decisions under AI assistance from a group of pathology doctors (pathologists). Two metrics were used to evaluate the appropriateness of AI reliance: Relative AI Reliance (RAIR) and Relative Self-Reliance (RSR). Results showed that even with groups of three pathologists, majority-voted decisions significantly increased both RAIR and RSR – by approximately 9% and 31%, respectively – compared to decisions made by one pathologist collaborating with AI. This increased appropriateness resulted in better precision and recall in the detection of mitoses. While our study is centered on pathology, we believe these insights can be extended to general high-stakes decision-making processes involving similar visual tasks.

1. Introduction

Although J.C.R. Licklider introduced the concept of 'man-computer symbiosis' in 1960 (Licklider, 1960), it was not until the last decade that this vision became a more promising reality (Jordan and Mitchell, 2015). By 2023, Artificial Intelligence (AI) has been increasingly discussed to augment humans in critical tasks (Bi et al., 2019; Surden, 2019; Grigorescu et al., 2020). Especially in the medical domain of pathology, AI has been showcased to increase doctors' accuracy and speed (Litjens et al., 2016; Lindvall et al., 2021; Van der Laak et al., 2021; Ba et al., 2022), consistency (Balkenhol et al., 2019; Van Bergeijk et al., 2023), and confidence (Gu et al., 2023b). However, because pathology AI was often trained from a limited dataset its performance varied while being applied to data from new patients and hospitals (Stacke et al., 2020; Aubreville et al., 2021; Gu et al., 2021). As such, it is critical for pathologists to develop appropriate reliance while collaborating with AI, i.e., to appropriately accept correct AI recommendations and reject the wrong ones.

Although there is a lack of data in pathology, research in the general domain has explored methodologies to develop appropriate reliance, focusing on reducing humans' over-reliance on AI (i.e., enhancing humans' ability to reject wrong AI recommendations). Strate-

* Corresponding authors.
E-mail addresses: ghy@ucla.edu (H. Gu), mhaeri@kumc.edu (M. Haeri).

https://doi.org/10.1016/j.ijhcs.2024.103315
Received 14 December 2023; Received in revised form 27 May 2024; Accepted 7 June 2024
Available online 21 June 2024

gies, including the cognitive forcing function (Buçinca et al., 2021) and altering the interaction speed (Park et al., 2019; Rastogi et al., 2022; Lebedeva et al., 2023), have shown promising results. Additionally, effective onboarding (Passi and Vorvoreanu, 2022) and improving AI literacy (Long and Magerko, 2020) were recommended and can be achieved by informing users of AI details (Cai et al., 2019b; Jacobs et al., 2021a,b). However, incorporating these methods into routine medical practice presents challenges: Cognitive forcing functions could drive medical practitioners to develop algorithm aversion (Efendić et al., 2020; Fogliato et al., 2022), which may cause them to dismiss AI recommendations even when they were correct. Moreover, previous studies have reported that the improvements in task accuracy with enhanced AI literacy were marginal (Lai et al., 2020; Leichtmann et al., 2023).

Another popular approach aims to employ explainable AI (XAI) to reduce over-reliance (Bussone et al., 2015; Lai et al., 2020; Zhang et al., 2020; Bansal et al., 2021). However, the efficacy of XAI is countered in part by the cognitive effort for understanding these explanations (Vasconcelos et al., 2023). Adding XAI-related content might increase doctors' cognitive burden, possibly causing them to overlook XAI. Therefore, there remains a pressing need for alternative strategies to foster appropriate AI reliance in medical applications.

By reviewing pathologists' decision-making workflows, we found that the critical decisions were usually determined through a combined judgment among multiple doctors (Black et al., 1999). The underlying intuition was that a group of pathologists might produce safer and more rational judgments while working together (Black et al., 1999). In the context of AI, recent studies have employed majority voting among pathologists' AI-assisted decisions to collect annotations for datasets (Bertram et al., 2019; Aubreville et al., 2020). However, there is a lack of empirical evidence supporting that such a majority voting approach would enable appropriate reliance.

This research aims to provide the validation of the majority voting on enabling the appropriate AI reliance in pathology decision-making, with a focus on a visual search task of detecting "mitosis", a critical histology pattern for tumor grading (Collan et al., 1996; Meyer et al., 2005). 32 medical professionals in pathology from ten institutions participated in a multi-stage user study, where they detected mitoses manually, first, and with AI assistance after a wash-out period. Here, the majority voting decisions were synthesized according to the AI-assisted decisions from an odd number of randomly-selected pathologist participants. Two metrics were employed to measure the appropriateness of AI reliance: "relative AI reliance" and "relative self-reliance" (Schemmer et al., 2023). The result showed that the majority voting decisions from as few as three pathologists showed significantly higher relative AI reliance (~9% increase) and relative self-reliance (~31% increase), compared to one pathologist collaborating with AI, respectively. The precision and recall of majority voting decisions also increased: Those from three AI-assisted pathologists could achieve a mean precision of 0.902 and a recall of 0.843. As a comparison, the mean precision and recall for one-pathologist-AI collaboration were 0.824 and 0.817, respectively. Furthermore, the majority voting decisions could also have a higher chance of achieving super-AI performance in the recall.

1.1. Contributions

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

2. Related work

2.1. Enabling appropriate AI reliance

According to Passi and Vorvoreanu (2022), Vasconcelos et al. (2023) and Schemmer et al. (2023), two goals should be achieved to enable appropriate AI reliance: (1) mitigating over-reliance, where humans can identify and reject AI's incorrect recommendations, and (2) reducing under-reliance, where humans can overcome their aversion of AI and accept its correct recommendations.

In the context of enabling appropriate AI reliance, this is a tendency in research to study mechanisms and counter-measures for over-reliance. For instance, the cognitive forcing function, which prompts users to think analytically before decision-making, has shown promise (Buçinca et al., 2021). Similarly, altering the interaction speed, where enlonging the AI response time, can instigate users' reflective thinking. Therefore, over-reliance incidents could be reduced (Park et al., 2019; Rastogi et al., 2022; Lebedeva et al., 2023). Other approaches aim to enhance users' onboarding process, such as improving AI literacy (Lai et al., 2020; Long and Magerko, 2020; Leichtmann et al., 2023), where users are informed of AI details (Cai et al., 2019b; Jacobs et al., 2021a,b). However, translating these approaches to the medical domain may encounter two challenges. Firstly, introducing cognitive forcing functions or altering interaction speed could develop 'algorithm aversion,' especially when medical tasks are time-sensitive (Efendić et al., 2020; Fogliato et al., 2022). Secondly, the efficacy of enhancing AI literacy also appeared marginal, possibly because of the difficulties in educating users within a limited timeframe (Lai et al., 2020; Leichtmann et al., 2023).

Besides these, another popular approach is XAI, aiming to reduce over-reliance by enabling users to understand AI's reasoning (Bussone et al., 2015; Lai et al., 2020; Zhang et al., 2020; Bansal et al., 2021). Nonetheless, numerous studies have failed to observe the anticipated effectiveness of XAI (Schemmer et al., 2022): The potential benefits of XAI may be offset by the cognitive efforts of interpreting them (Vasconcelos et al., 2023). Given the already high cognitive demands of medical professionals, this might result in XAI being less referred to, countering its potential benefits. This issue of appropriateness usage of XAI in medicine was raised by Holzinger et al. (2019). Further research suggested causability, an ability of an explanation that can enable casual understanding of medical experts, should also considered and measured to achieve better efficiency, effectiveness, and user satisfaction (Holzinger et al., 2020; Plass et al., 2023).

Notably, most of the research mentioned above focuses on scenarios where one human collaborates with AI. Different from these studies, our approach learns from how critical pathology decisions are usually made — through a group of pathologists (Black et al., 1999). Specifically, we employ a majority voting approach to synthesize decisions from multiple AI-assisted pathologists. Based on the results, the majority voting approach can effectively enable appropriate AI reliance, which sheds light on the potential of involving multiple professionals in critical decision-making.

2.2. Decision-making processes by multiple medical professionals

Different from one medical professional examining the specimens (Pohn et al., 2019; Pena and Andrade-Filho, 2009), decision-making processes by medical professionals require communication, discussion, and result sharing (Murphy et al., 1998; Black et al., 1999). Nowadays, there are three primary approaches: (1) the Delphi method, (2) the nominal group technique, and (3) the consensus development conference. These methods have been historically utilized in medical decision-making and guideline formulation (Murphy et al., 1998; Black et al., 1999).

The Delphi method follows an iterative process: Each group member makes a decision first anonymously. Next, their opinions are collected,

summarized, and then sent back to all members. Upon reviewing this summary, each member may choose to modify their opinions secretly. This process may be iterated multiple times to resolve potential conflicts (Taze et al., 2022). The nominal group technique (Van de Ven and Delbecq, 1972) also starts by collecting each member's opinions. Then, in a structured face-to-face meeting, these opinions are presented and discussed. Next, each member ranks the presented opinions according to their preferences. These rankings will be summed and posted for further discussion (McMillan et al., 2016). The consensus development conference (Ferguson, 1996) is more open in its structure. Group members are presented with evidence by external experts during a series of face-to-face meetings. Group members can then question the expert presenters, and attempt to reach an agreement afterward.

Regarding employing multiple pathologists to make decisions with AI assistance, recent research has implemented majority voting among three doctors to label data for AI development (Bertram et al., 2019; Aubreville et al., 2020). This process involves two pathologists independently annotating data with AI assistance. If there were conflicts, a third pathologist joined and annotated again to formulate majority (Montezuma et al., 2023). However, these studies primarily focus on data labeling for AI development, often featuring non-diverse participant pools (typically three pathologists) with similar backgrounds. Furthermore, they lack analyses of AI reliance metrics, which is critical for assessing the method's quality.

Our study aims to fill the gap by providing a comprehensive experiment and evaluation of the majority voting on pathology decision-making. To achieve this, we hosted a multi-stage, multi-institutional user study involving 32 medical professionals in pathology with varied experience levels. We examined the quality of these majority-voted, AI-assisted decisions from two angles: reliance on AI and correctness. Additionally, we evaluated the potential costs of employing this method, providing empirical data for future research in HCI, cognitive science, and medicine.

2.3. Human-AI collaboration in pathology

AI, particularly deep learning, holds promise in performing a wide variety of pathology tasks to assist medical professionals (Regitnig et al., 2020), ranging from conducting high-level diagnoses (e.g., prostate cancer grading (Pantanowitz et al., 2020)), to detecting low-level pathological patterns (e.g., cell detection (Amgad et al., 2022)). Notably, recent studies have suggested that AI's performance is on par with human experts in specific pathology tasks (Hekler et al., 2019; Zhang et al., 2019; Wang et al., 2023). Due to legislative and ethical concerns, AI algorithms and software cannot replace pathologists' examinations (Chauhan and Gullapalli, 2021; Veale and Zuiderveen Borgesius, 2021). Instead, they are regarded as medical devices to assist doctors,[1] with one designed for prostate cancer pathology receiving the first official approval in 2021.[2] By 2023, a plethora of human-AI collaborative tools have been introduced, demonstrating improvements in speed and correctness in detecting pathological patterns (Litjens et al., 2016; Lindvall et al., 2021; Van der Laak et al., 2021; Ba et al., 2022), inter-observer consistencies (Balkenhol et al., 2019; Van Bergeijk et al., 2023), mental workload and confidence (Gu et al., 2023b,c), compared to pathologists examining manually.

Among these improvements, of particular interest is achieving complementary team performance, where pathologist-AI collaboration could outperform both the pathologist and AI (Bansal et al., 2021): In 2016, Wang et al. reported that combining AI and pathologist's predictions could reduce error rates in breast cancer classification, theoretically confirming the existence of such complementary team performance (Wang et al., 2016). Despite this theoretical backing, there is a lack of empirical evidence to support it for the pathology domain. In the general domain, several studies have failed to observe the task accuracy improvement in human-AI collaboration compared to AI alone (Bansal et al., 2021; Lai and Tan, 2019; Jacobs et al., 2021b). This issue may stem from users' accepting incorrect AI recommendations, a factor that can significantly impact the outcome of human-AI collaboration (Kaur et al., 2020; Cao and Huang, 2022; Vasconcelos et al., 2023).

To date, research remains sparse on how pathologists would rely on AI. This work fills this gap and by recruiting pathology professionals and studying their AI reliance, which can help future researchers understand pathologists' behavior, and develop potential solutions to enable appropriate AI reliance.

3. Task design & medical background

3.1. Task selection & generalizability of the task

This work selects the task of mitosis (a type of histology pattern) detection in brain tumors of meningiomas (Fig. 1(a)). The significance of mitosis stems from its critical role in tumor assessment and patient management for meningiomas (Cree et al., 2021; Louis et al., 2021; Goldbrunner et al., 2021). Despite their importance, pathologists' evaluation of mitoses often faces substantial difficulties. The intricacies lie in mitotic figures' small size, low prevalence, and heterogeneous distribution (Aubreville et al., 2020; Bertram et al., 2020). These complexities contribute to low reported sensitivities, consistencies among pathologists, and examination efficiencies for mitosis evaluation (Collan et al., 1996; Meyer et al., 2005; Veta et al., 2016; Gu et al., 2023c), which could negatively impact medical outcomes.

According to the 2021 World Health Organization central nervous system tumor classification guidelines, mitosis serves as a critical diagnostic criterion for grading numerous brain tumors, such as IDH-mutant astrocytoma, oligodendroglioma, and ependymoma (Louis et al., 2021). Going beyond mitoses, pathologists may also be required to detect small-scale, sparsely distributed patterns in large scans, such as finding small tumor deposits within lymph nodes in breast cancer or malignant melanoma (Regitnig et al., 2020). In a more general context, similar visual search tasks also exist in high-stakes domains where AI assistance could be valuable. For instance, security personnel must swiftly identify potential threats like explosives in X-ray scans (Wolfe et al., 2007), and emergency responders rely on timely assessments of disaster impacts from satellite imagery (Morrison et al., 2023).

3.2. Sample selection & mitosis ground truth acquisition

Meningioma specimens were collected from a local hospital after receiving ethics approval. These specimens were digitized into 19 digital slides with an Aperio CS2 Scanner (Manufacture: Leica, Germany). A specialist pathologist examined these slides and selected 51 regions of interest (ROIs) based on predefined criteria. Each ROI has a dimension of 1600×1600 pixels (400×400 μm), with one example shown in Fig. 1(a). This image dimension matches the field-of-view under the 40× objective lens in light microscopy, which can reduce the mental effort for pathologists to adapt to the digital interface.

As for collecting the mitosis ground truth, two residents independently annotated all 51 images initially. Next, a third specialist pathologist reviewed these initial annotations and provided a final decision. To ensure the accuracy of the ground truth, the three doctors referred to the results of an additional antibody test (the Phosphohistone-H3 immunohistochemistry test, a mitosis indicator usually used in medical research (Duregon et al., 2015; Fukushima et al., 2009), Fig. 1(b)) in the ground truth annotation process.

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

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

[2] https://www.fda.gov/news-events/press-announcements/fda-authorizes-software-can-help-identify-prostate-cancer.

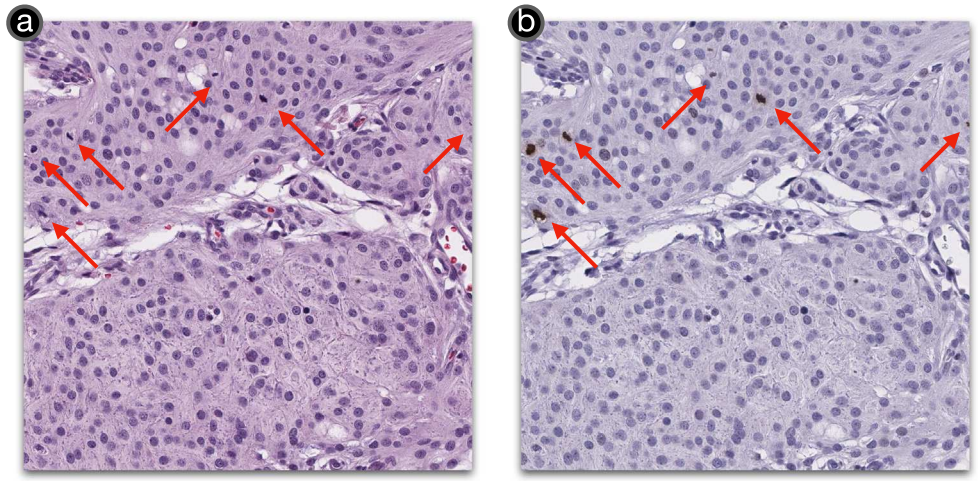

Fig. 1. (a) An example region-of-interest image used in the user study, with arrows pointing at the ground truth mitoses; (b) The anti-body test used by the three doctors to annotate the ground truth mitoses. Mitoses were shown in brown (as pointed by the arrows) in the anti-body test.

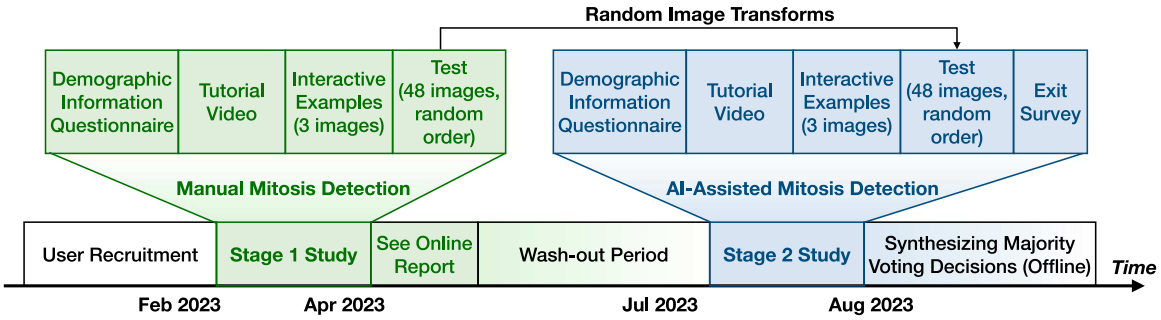

Fig. 2. Organization of the user study.

3.3. Experience level of pathologists

In the United States, pathology professionals can be classified into four levels based on their training progress and experience (Genzen, 2013):

1. A **medical student** is currently receiving medical education.
2. A **resident** has earned their Medical Doctor or an equivalent degree and is in post-graduate residency training.
3. A **general pathologist** has completed their residency training and holds general board certification in pathology.
4. A **specialist pathologist** has received/ is undergoing further training in a sub-specialty area (in this study, neuropathology) after becoming certified as a general pathologist.

Regarding familiarity with the mitosis detection task, specialist pathologists are expected to have the highest level because of their sub-specialty training. General pathologists should have a moderate familiarity, having acquired their general board certification. As for residents and medical students, their familiarity depends on their exposure during rotations and any subsequent training they have received. In this study, 32 medical professionals from ten institutions participated, covering all four aforementioned categories.

4. User study

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

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

4.1. Participants

Participants were recruited through sending emails to the mailing list and snowball recruitment. As a result, 32 pathology professionals from 10 medical centers in the United States took part in both study stages, including 12 specialist pathologists, six general pathologists, ten residents, and four medical students.[3] The demographic information of participants is shown in the supplemental material.

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

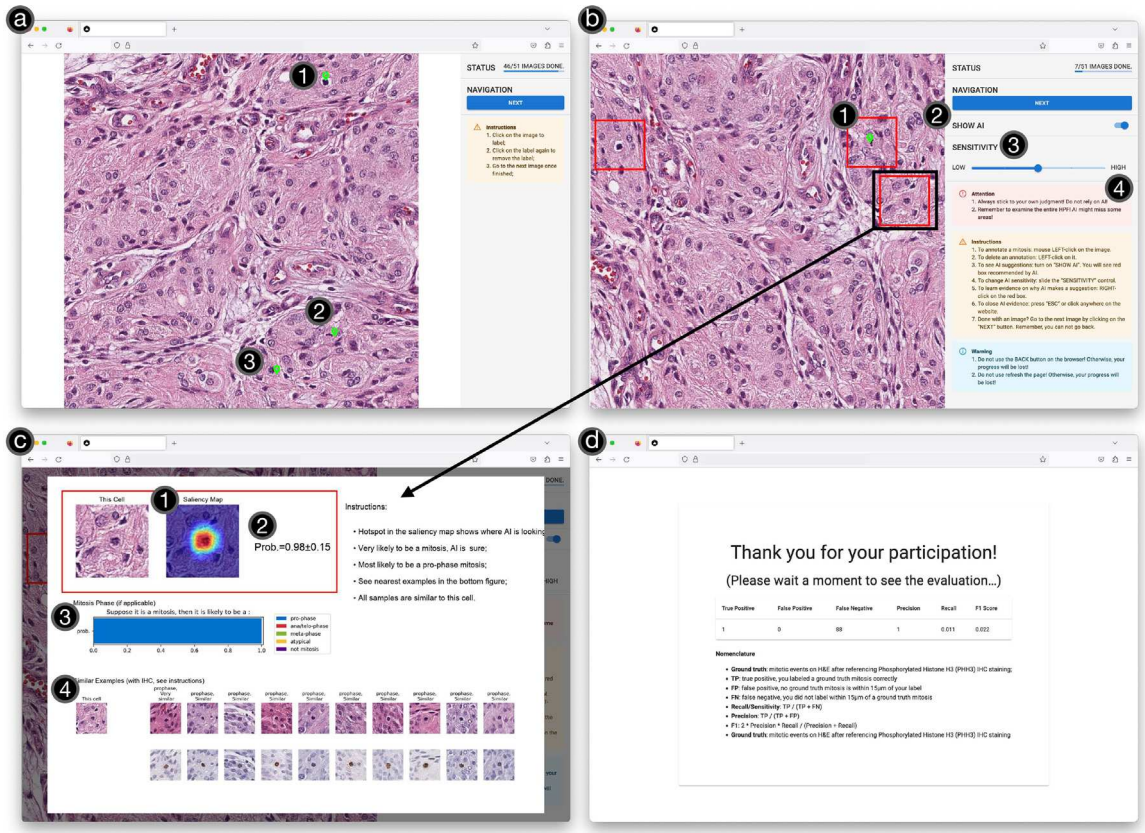

Fig. 3. Screenshots of the mitosis study websites: (a) The manual mitosis detection website in the stage 1 study. The user could left-click on the image to leave a mark for each mitosis detected (①–③). (b) The AI-assisted mitosis detection website in the stage 2 study. The interface added ① the AI recommendation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

4.2. Study procedure

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

1. **Demographic information**: Participants filled in a demographic information questionnaire.
2. **Tutorial**: Participants saw a tutorial video describing how to participate, followed by an interactive tutorial of three example images. No AI details were revealed to participants.
3. **Test**: Participants examined the 48 images without (stage 1) or with (stage 2) AI assistance. Their task was to detect and report mitoses from these images with their threshold of daily practice.

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

- **Random image transforms**: Including random flipping (vertical and/or horizontal) and random rotation (randomly chosen from {0°, 90°, 180°, and 270°}). For instance, the image shown in Fig. 3(b) was rotated 270° anti-clockwise from that in Fig. 3(a).
- **Wash-out period and ground truth blinding**: After completing stage 1, participants received personalized online report documents highlighting disagreements between their mitosis reportings and the ground truth. After two weeks, they were prevented from accessing these online documents. Next, after a wash-out period of three months, they were invited to participate in the stage 2 study.

4.3. User interfaces & key features

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

4.3.1. Stage 1: Manual mitosis detection

This interface only showed participants the images and logged their interactions (Fig. 3(a)). If the user found a mitosis, they could left-click on where it resided to leave a mark (Fig. 3(a)①–③). The user could go to the next image after examining one. However, they could not return to the previous image to ensure a precise measurement for time consumption. After all images were examined, a status page (Fig. 3(d)) was displayed to inform the participant of the performance of their mitosis detection.

4.3.2. Stage 2: AI-assisted mitosis detection

The AI model used in this stage was an EfficientNet-b3 Convolutional Neural Network (CNN), trained from a meningioma mitosis dataset (Tan and Le, 2019; Gu et al., 2023a). The website displayed AI mitosis detections through recommendation boxes (Fig. 3(b)). Additionally, following previous works, we included four components to mitigate the negative influence of improper AI reliance:

- **Warning messages**: A "black-box" style[4] warning message was presented in the tutorial video, suggesting that the users should

[4] ... the highest safety-related warning assigned by the U.S. Food and Drug Administration (Delong and Preuss, 2019).

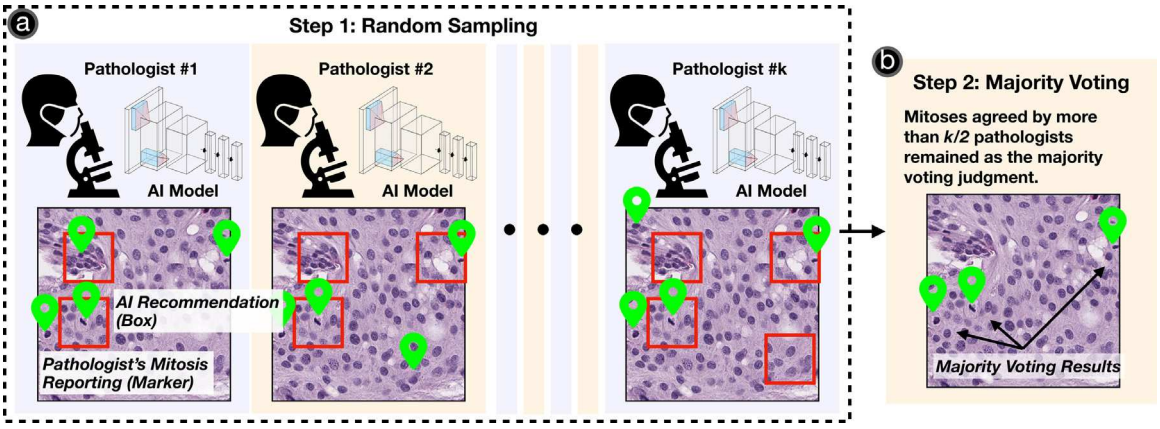

Fig. 4. Steps for synthesizing the majority voting decisions from k AI-assisted pathologists: (a) random sampling: mitosis reportings from an odd number of k randomly-sampled, AI-assisted pathologists were collected, (b) majority voting: mitoses candidates reported by $> k/2$ pathologists remained as the final decision.

always rely on their judgments. The message was also shown in a highlighted box on the website (Fig. 3(b)④).

- **XAI**: Each AI recommendation was accompanied by an evidence card which attempts to provide XAI assistance (Plass et al., 2023). The user could right-click on the AI recommendation box to see the XAI evidence card *on-demand*. Four popular XAI techniques were included following previous work (Evans et al., 2022), including:
  - **Saliency map**: Generated by GradCAM++ (Chattopadhay et al., 2018).
  - **Confidence level**: Including a probability score and a trust score (Zhang et al., 2020). The trust score was the geometric mean of noise (Ayhan and Berens, 2018) and random AI variances (Gal and Ghahramani, 2016) of the AI prediction.
  - **Subclass**: A bar plot showcasing potential subclasses of the mitosis (i.e., pro-phase, meta-phase, ana/telo-phase, atypical, and not mitosis) in this AI recommendation.
  - **Similar examples**: A set of ten similar instances was retrieved from an annotated dataset that includes paired Hematoxylin and Eosin — immunohistochemistry staining (Cai et al., 2019a).

  Counterfactual explanations were not used because of the low quality of the retrieval results achieved by the our AI model.
- **Personalized AI adjustments**: The user could toggle on/off AI recommendations by interacting with the "Show AI" switch (Fig. 3(b)②) (i.e., AI on-request, suggested by Gaube et al. (2021)) and adjust the AI sensitivity (Fig. 3(b)③) according to their preferences. The website provided five AI sensitivity settings for users: "lowest", 'low", "medium", "high", and "highest". A higher sensitivity would include more AI recommendations with lower probabilities.
- **Random image order**: The 48 images were presented to participants in a random order to prevent users from anchoring on AI based on their initial impressions (i.e., the ordering effect Nourani et al., 2021).

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

4.4. Synthesizing majority voting decisions from groups of AI-assisted participants

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

Step 1 **Random Sampling**: Mitosis reportings from an odd number k participants from stage 2 were aggregated as a group (Fig. 4(a)). Members in a group were sampled randomly from the participant pool without replacement.

Step 2 **Majority Voting**: Mitoses candidates reported by more than half of members ($k/2$) in the group remained as the final majority voting decision (Fig. 4(b)).

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

4.5. Measures & statistics

4.5.1. Utilization of AI & XAI (RQ1)

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

- **AI activation rate**: Indicating the percentage of the 48 test images where the AI was activated at least once (Eq. (1)).
- **AI active time percentage**: Since the participant might deactivate the "Show AI" feature, this metric represents the percentage of time when the "Show AI" feature stayed active during the entire stage 2 study (Eq. (2)).

$$\text{AI activation rate} = \frac{\sum_{i=1}^{48} \mathbb{1}[\text{"Show AI" in image}_i == \text{"On"}]}{48} \times 100\% \quad (1)$$

$$\text{AI active time percentage} = \frac{\sum_{i=1}^{48} T[\text{"Show AI" in image}_i == \text{"On"}]}{\sum_{i=1}^{48} \text{Time consumption on image}_i} \times 100\% \quad (2)$$

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

- **XAI activation rate** was calculated according to Eq. (3). The number of "AI recommendations in image$_i$" was counted based on the highest sensitivity set by a participant while they examined the image$_i$. If the "Show AI" was not toggled on in an image, then it was not counted.
- **XAI activation time** was measured by the time elapsed between a participant opening and closing an XAI evidence card.

$$\text{XAI activation rate} = \frac{\sum_{i=1}^{48} |\text{XAI opened in image}_i|}{\sum_{i=1}^{48} |\text{AI recommendations in image}_i| \times \mathbb{1}[\text{"Show AI"} == \text{"On"}]} \times 100\% \quad (3)$$

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

Fig. 5. Combinatorics for reliance incidents in the condition of one pathologist collaborating with AI (i.e., one-human-AI) for the mitosis detection task. This chart is adopted from the framework described in Schemmer et al. (2023).

4.5.2. Reliance on AI (RQ2)

We used the categorization proposed by Schemmer et al. (2023) to define the incidents related to the reliance. Four types of events were defined under the categorization: (1) correct self-reliance, (2) incorrect AI reliance (over-reliance), (3) correct AI reliance, and (4) incorrect self-reliance (under-reliance). The criteria for judging these events were based on the true-positive (TP), true-negative (TN), false-positive (FP), and false-negative (FN) detections.[5] We adopted the framework in Schemmer et al. (2023) for the mitosis detection task, which is summarized in Fig. 5.

Schemmer et al. further introduced two normalized metrics, Relative AI Reliance (RAIR), and Relative Self-Reliance (RSR), to represent the Appropriateness of Reliance (AoR). The RAIR relates to the under-reliance events (Eq. (4)). And the RSR relates to the over-reliance events (Eq. (5)). The Appropriateness of Reliance is encapsulated by the tuple of PAIR and RSR (Eq. (6)), which can be graphically represented on a 2D chart with the RAIR on the x-axis and the RSR on the y-axis.

$$\text{Relative AI reliance (RAIR)} = \frac{\text{Correct AI Reliance}}{\text{Correct AI Reliance} + \text{Under-reliance}} \tag{4}$$

$$\text{Relative Self reliance (RSR)} = \frac{\text{Correct Self Reliance}}{\text{Correct Self Reliance} + \text{Over-reliance}} \tag{5}$$

$$\text{Appropriateness of Reliance (AoR)} = (RSR; RAIR) \tag{6}$$

To measure AI reliance on majority voting decisions, we also implemented the majority voting process for stage 1. To ensure a "within-subject" nature of the analysis, for each majority voting run for stage 2, a vis-à-vis majority voting from the same group of participants in stage 1 was conducted. The definitions of "human decisions", "AI recommendations", and "human-AI decisions" were adjusted to fit the majority voting condition and are summarized in Table 1.

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

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

Table 1
Modified definitions to measure AI reliance for the majority voting decisions synthesized from a group of k pathologists.

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

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

4.5.3. Correctness of mitosis detection (RQ3)

We used *precision* (Eq. (7)) and *recall* (Eq. (8)) to measure the correctness of the mitosis detection.

$$Precision = \frac{TP}{TP + FP} \tag{7}$$

$$Recall = \frac{TP}{TP + FN} \tag{8}$$

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

Because our previous work showed AI achieved higher overall performance than all participants in stage 1 (Gu et al., 2024), the "complementary team performance" in this work refers explicitly to cases where the human+AI approach outperforms AI (**RQ3**, i.e., super-AI performance). Here, the AI operating point was selected based on the best threshold in the model validation process. For precision and recall, we defined the "success rate of achieving super-AI performance" using Eq. (9). This equation was applied to both the one-human-AI collaboration, and majority voting decision conditions with group sizes k ranging from 3 to 27.

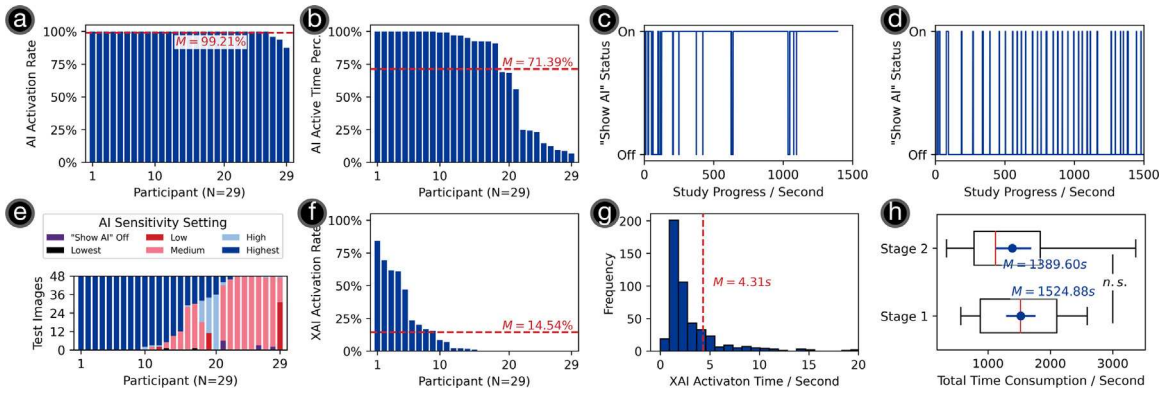

Fig. 6. (a) Bar-plot of AI activation rates; (b) Bar-plot of AI active time percentage; Example plots showing how "Show AI" status changed for (c) a participant with a high (92.31%) AI active time percentage and (d) a participant with a low (14.48%) AI active time percentage; (e) Stacked bar-plot of participants' AI sensitivity settings; (f) XAI activation rates; (g) Histogram of XAI activation time; (h) Box-whisker plot of total time consumption of each participant spent on image examination in the stage 1 and stage 2 study. No significance (n.s.) was observed between the two stages.

$$\text{Success Rate} = \frac{\text{Number of participants/runs exceeding AI performance}}{\text{Total number of participants/runs}} \times 100\% \quad (9)$$

5. Result

3/32 participants in stage 2 chose not to activate AI recommendations at all for over 45/48 test images. Therefore, they were classified as non-AI users and were excluded from subsequent analyses. For the remaining 29 participants, we report the utilization of AI and XAI in Section 5.1. 25/29 of the participants majorly set the sensitivity as either "highest" (N = 15) or "medium" (N = 10) during the stage 2 study, and they were included in the AI reliance analysis (Section 5.2). The responses from all 29 AI-users were used for correctness analyses in Section 5.3.

5.1. Utilization of AI & XAI

The mean AI activation rate was $M = 99.21\%$ ($SD = 0.481\%$, $CI_{95} = [98.13\%, 100.00\%]$, Fig. 6(a)).[6] And the mean AI active time percentage was $M = 71.39\%$ ($SD = 6.713\%$, $CI_{95} = [57.75\%, 83.94\%]$, Fig. 6(b)). 21/29 participants had > 50% AI active time percentages, with an example of how they interacted with the "Show AI" feature shown in Fig. 6(c), which suggests the user kept the AI activated for the majority of the time, with occasional brief flickering between turning it off and on during the initial interactions. The remaining 8/29 participants had < 25% AI active time percentages: Although the "Show AI" feature was majority deactivated, these participants would still activate AI recommendations briefly while examining each image (Fig. 6(d)). Interestingly, this pattern matches the cognitive forcing function (Buçinca et al., 2021) although these participants had not been instructed to do so.

For AI sensitivity settings, 15/29 participants set for the "highest" for over half of the ROI images. The remaining participants preferred to set the AI sensitivity as "high" (1/29), "medium" (10/29), "low" (1/29), or showed no clear preference (2/29), as shown in Fig. 6(e).

Regarding XAI utilization, the mean XAI activation rate was $M = 14.54\%$ ($SD = 4.537\%$, $CI_{95} = [6.43\%, 24.25\%]$, Fig. 6(f)). Specifically, 4/29 participants had XAI activation rates higher than 50%, while 14/29 participants did not activate any XAI at all. The mean XAI activation time was $M = 4.31$ seconds ($SD = 0.719$ s, $CI_{95} = [3.17\text{ s}, 5.95\text{ s}]$, Fig. 6(g)).

On average, participants spent 25 min and 25 s examining all 48 test images in stage 1, and 23 min and 9 s in stage 2 (Fig. 6(h)). The total time consumption did not show a significant difference between the two stages (Wilcoxon rank-sum test, $p = 0.31$).

5.2. Reliance on AI

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

The mean RSR of one-human-AI collaboration was $M = 0.735$ ($SD = 0.037$, $CI_{95} = [0.657, 0.803]$, see Fig. 7(b)). As a comparison, the mean RSR of majority voting decisions of $k = 3$ was $M = 0.964$ ($SD = 0.003$, $CI_{95} = [0.959, 0.970]$). The RSR of majority voting decisions for $k = 5, 7, 9$ were 0.968, 0.976, and 0.967. Similarly, all four majority voting conditions led to higher RSR (~31% increase). A Kruskal–Wallis test showed a significant difference among the RSR values across five conditions ($\eta^2_H = 0.178$, $p < 0.001$). Post-hoc Dunn-Bonferroni test showed significance in comparison pairs of one-human-AI vs. majority voting decision from group sizes of $k = 3$ ($p < 0.001$), $k = 5$ ($p < 0.001$), $k = 7$ ($p < 0.001$), and $k = 9$ ($p < 0.001$).

Fig. 7(c) presents the appropriateness of reliance (AoR) scatter plots for five conditions. These plots demonstrate that majority voting decisions could improve higher RAIR and RSR simultaneously, indicating a high level of AoR was achieved.

5.3. Correctness of mitosis detection

As shown in Fig. 8(a), the mean precision of one-human-AI collaboration was $M = 0.824$ ($SD = 0.023$, $CI_{95} = [0.776, 0.867]$). For majority voting decisions of $k = 3$, the mean precision was $M = 0.902$ ($SD = 0.004$, $CI_{95} = [0.893, 0.910]$). The majority voting of $k = 5, 7, 9$ had mean precisions of 0.924, 0.934, and 0.934, respectively. The AI achieved a higher precision of 0.961. All four majority

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

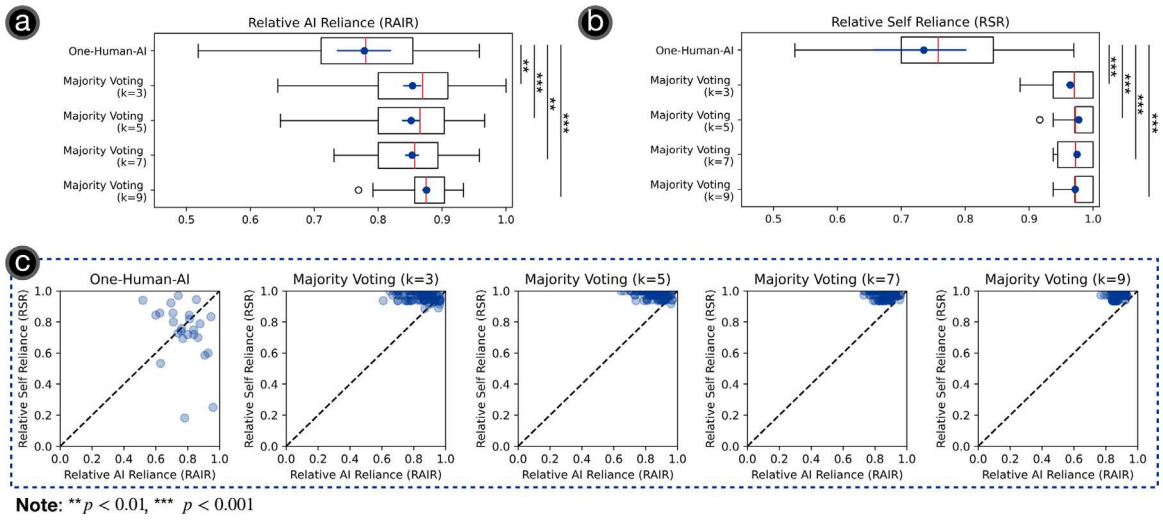

Fig. 7. Box-whisker plots of (a) RAIR and (b) RSR for the five conditions of one-human-AI collaboration, and majority voting decisions ($k = 3, 5, 7, 9$); (c) Scatter plots for appropriateness of reliance for these five conditions.

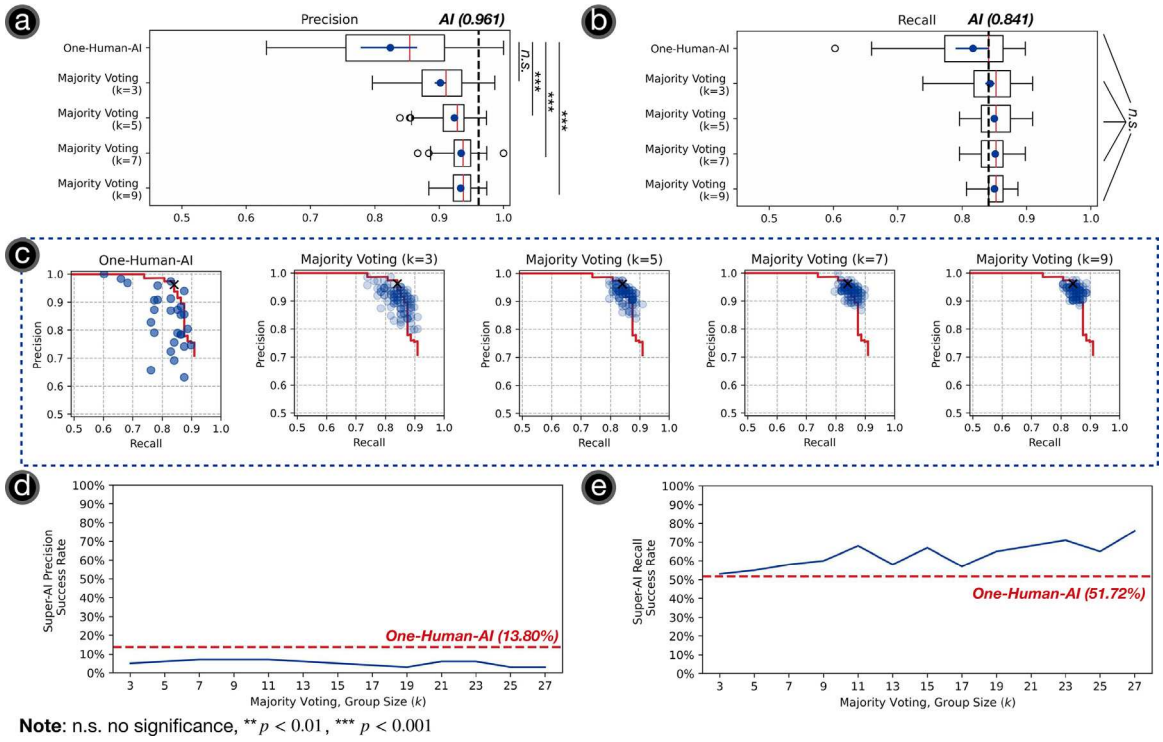

Fig. 8. Box-whisker plots of precision and recall for the five conditions of one-human-AI collaboration, and majority voting decisions ($k = 3, 5, 7, 9$); (c) Precision–recall plots for mitosis detection for these five conditions. The red line represents the precision–recall curve of AI, and the 'x' marker indicates the AI's performance at a threshold determined by the best validation performance. The success rates of achieving super-AI performance (i.e., percentage of human+AI cases where their performance is higher than both humans and AI) for the criteria of (d) precision and (e) recall.

voting conditions achieved higher precision (~8% increase) than the one-human-AI collaboration. A Kruskal–Wallis test showed that the precision significantly differed across five conditions ($\eta^2_H = 0.150$, $p < 0.001$). Post-hoc Dunn-Bonferroni test did not observe significance in the comparison pair of one-human-AI vs. majority voting decision $k = 3$ ($p = 0.715$). Statistical significance was observed for comparison pairs of one-human-AI vs. majority voting decision $k = 5$ ($p < 0.001$), $k = 7$ ($p < 0.001$), and $k = 9$ ($p < 0.001$).

The mean recall of one-human-AI collaboration was $M = 0.817$ ($SD = 0.013$, $CI_{95} = [0.790, 0.841]$, see Fig. 8(b)). Majority voting decisions of $k = 3$ had a mean recall of $M = 0.843$ ($SD = 0.003$, $CI_{95} = [0.838, 0.851]$). Majority voting decisions of $k = 5, 7, 9$ had mean precisions of 0.850, 0.851, and 0.850. In comparison, AI achieved a precision of 0.841. Kruskal–Wallis test did not show that the recall differed significantly across five conditions ($\eta^2_H < 0.001$, $p = 0.774$).

Fig. 8(c) presents the precision–recall scatter plots for the five conditions. The plots reveal that the majority voting decision exhibits lower variation in both precision and recall compared to the one-human-AI collaboration, indicating a more robust performance. This observation is further supported by the lower SD values for the majority voting decisions, as reported above.

Regarding the success rates for achieving super-AI performance, for precision, none of the majority voting conditions (i.e., $k = 3 \rightarrow 27$) was higher than the success rate achieved by one-human-AI collaboration

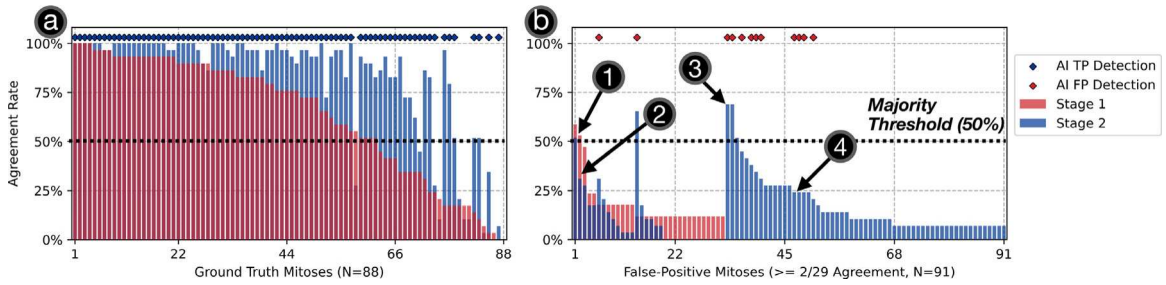

Fig. 9. Bar plots for the agreements among 29 participants for (a) 88 ground-truth mitoses, and (b) 91 false-positive mitoses that at least two participants agreed on. The diamond markers (◊) stand for the AI detections under the "Highest" AI sensitivity setting. ① An example of under-reliance that might not be addressed by the majority voting; ② An example of under-reliance that might be addressed by the majority voting; ③ An example of over-reliance that might not be addressed by the majority voting; and ④ An example of over-reliance that might be addressed by the majority voting.

(13.87% success rate, Fig. 8(d)). On the other hand, for recall, all majority voting conditions had higher success rates compared to one-human-AI collaboration (51.72% success rate): As shown in Fig. 8(e), the lowest success rate was observed at $k = 3$ (53% success rate), and the highest was achieved at $k = 27$, reaching 76%.

6. Discussion

6.1. Summary of result

6.1.1. Summary of RQ1

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

The utilization of XAI was relatively low; only four participants opened more than 50% of the XAI evidence, while nearly half of the participants did not open any. Even when XAI was opened, the time spent by participants on viewing XAI was relatively short (about four seconds) – in the context of pathologist-AI collaboration, the effectiveness of XAI in mitigating over-reliance may be limited. This is likely because the time-pressing nature of the pathology task outweighed the benefit of XAI explanations, causing pathologists to use XAI less in practice. In light of this, we argue that alternative approaches, such as the majority voting used in this study, need to be investigated to enable appropriate AI reliance for future pathology applications.

6.1.2. Summary of RQ2

Pair-wise statistical tests revealed significant improvements in both RAIR and RSR metrics for majority voting decisions ($k = 3, 5, 7, 9$), compared to one pathologist collaborating with AI. Specifically, RAIR showed an approximate 9% increase, and RSR showed about 31% increase. The PAIR-RSR scatter plots indicated simultaneous improvements in both metrics. Such results demonstrate a reduction in the proportion of over-reliance against correct self-reliance events, and under-reliance against correct AI reliance events, indicating a higher level of appropriateness of reliance was achieved (according to the definitions in Schemmer et al. (2023)).

6.1.3. Summary of RQ3

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

All majority voting conditions for $k = 3 \rightarrow 27$ did not show a higher success rate in achieving super-AI precision than one-human-AI collaboration. This is because AI had a high precision of 0.961, and there was a lack of space for improvement. For recall, all majority voting conditions ($k = 3 \rightarrow 27$) showed higher success rates. Notably, the highest success rate, 76%, was achieved at $k = 27$, indicating a 46.95% increase over the one-human-AI collaboration condition (51.72% success rate).

6.2. The mechanism and cost of majority voting

To further explore why the majority voting mechanism was effective, we introduced a metric, "agreement rate", defined as the percentage of participants the reported a cell as a mitosis (regardless of its actual status). We calculated the agreement rates of the 29 participants in both stage 1 and stage 2 studies. These agreement rates covered all 88 ground truth mitoses (Fig. 9(a)) and 91 false-positive mitoses reported by at least two participants (Fig. 9(b)). According to Section 4.4, cells with agreement rates higher than 50% should be kept as the majority voting decisions. While Fig. 9 is not directly applicable for interpreting results in smaller sub-groups (e.g., $k = 3$), it illustrates the general trends in participants' agreement rates when influenced by AI. The data revealed two key insights:

- **Reducing Over-Reliance on AI False Positives**: AI's false-positive detections led to higher agreement rates among participants (as shown in Fig. 9(b)③), suggesting participants' tendency of over-reliance in at stage 2. The majority of these false-positive detections did not achieve agreement rates higher than 50% (Fig. 9(b)④). In other words, from a group's perspective, it was not usual for the majority of participants to consistently over-rely when AI made false-positive mistakes. Therefore, the over-reliance can be reduced by the majority voting mechanism.
- **Reducing Under-Reliance on Human False Positives:** At stage 2, participants may make the same false-positive mistake as in stage 1, even when AI correctly suggested negative (Fig. 9(b)①). This suggests that the under-reliance incidents happened when one participant collaborated with AI. Nevertheless, agreement rates for these false positives rarely exceeded the 50% majority threshold (Fig. 9(b)②), indicating that majority voting could reduce under-reliance.

To understand the underlying cost of the majority voting mechanism, we analyzed time consumption spent on employing multiple pathologists, and its association with the correctness. Specifically, we conducted 100 majority voting runs for each group size (k) ranging from 3 to 27. We applied Pearson's correlation analysis to assess the relationship between precision or recall achieved in each run and its corresponding time consumption. We found a moderate positive correlation between precision and time consumption (Pearson's $r = 0.39$, $p < 0.001$, $N = 1300$, Fig. 10(a)), and a weak positive correlation between recall and time consumption (Pearson's $r = 0.14$, $p < 0.001$,

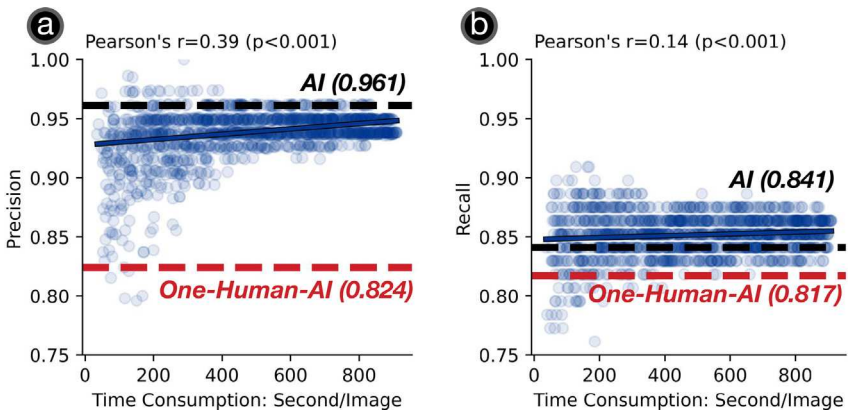

Fig. 10. Linear regression plots studying the relations between (a) precision-time consumption, and (b) recall-time consumption while synthesizing majority voting decisions, $k = 3 \rightarrow 27$, $n = 100$ for each k.

$N = 1300$, Fig. 10(b)). Certain runs with a relatively small time consumption could reach considerable precision and recall. Note that this is a 'bare minimum' estimation: Delays caused by coordinating pathologists should be taken into account in practical applications.

6.3. On developing structured decision-making processes with AI+k

Different from traditional one-human-AI collaboration (AI+1), this study sets the first step towards multiple medical professionals collaborating with AI (AI+k) using a simple majority voting technique. We argue that this majority voting approach has three advantages: (1) It is flexible and has a simple structure, eliminating the need for face-to-face or online discussions; (2) It keeps participants anonymous, thus reducing potential social pressure; (3) It is inherently democratic, ensuring that each participant's opinion has an equal weight. We found that this majority voting approach could effectively improve the appropriateness of reliance, and achieve higher-quality medical decisions. As for the limitations of majority voting, one may argue that this approach does not incorporate the discussion process, and decisions with conflicts (i.e., ~50% agreement rates) cannot be addressed easily.

Future works might explore AI+k decision-making techniques that involve structured or semi-structured face-to-face discussions (Black et al., 1999). Traditionally, these discussions were moderated by the humans. Nonetheless, we envision that future AI can not only help each group member to reach a decision (e.g., help pathologists detect mitoses in this study), but can moderate the discussions. For instance, a large language model (LLM) (Min et al., 2023) might anonymously gather and summarize comments from each group member and present a consolidated overview to the group. Members could then have an opportunity to revise their decisions after hearing from the LLM's summary. Given the LLM's omni-availability, no conflict of interest, and impartiality to authority or personal factors, such AI-facilitated discussions could offer advantages in speed and bias correction, compared to traditional discussion coordination with human moderators.

6.4. Towards efficient & reliable medical decisions with AI+k

Section 6.2 showed that the performance of majority voting decisions from AI+k showed a positive relation to the time consumption. In other words, in general, the more medical professionals involved, the higher the quality of the majority voting decision. Typically, high-risk medical decisions involve 7–10 group members (McMillan et al., 2016), while groups as large as 27 done in this study were quite rare. Therefore, considering the time taken to reach a result, we argue that not all medical decisions necessitate the AI+{*large k*} approach: cases with high confidence from both AI and humans could be adjudicated by smaller groups with as few as three experts, while those with low AI confidence or prone to human errors could benefit from incrementally larger group sizes, which can yield better and more robust outcomes.

Determining the optimal balance between decision-making and time expenditure has been well-explored in previous crowd-sourcing works (Daniel et al., 2018). However, one should be aware that the workflow of medical professionals is usually different from that of general users, and their preferences in using AI and XAI may also vary (as shown in Section 5.1). Therefore, future research should focus on exploring which AI+k methods can seamlessly integrate into the workflow of medical professionals, effectively balancing efficiency and reliability in the medical decisions of multiple doctors. Additionally, investigating the role of counterfactual explanations (Zhou et al., 2022; Del Ser et al., 2024) to build trust and facilitate appropriate AI reliance could complement approaches like majority voting, potentially improving interpretability and familiarity with the decision process when integrating AI into risk-sensitive medical workflows.

6.5. Limitations & future work

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

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

7. Conclusion

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

Funding resources

This work is funded partly by U.S. National Science Foundation (CAREER 2047297) to Xiang 'Anthony' Chen and the University of Kansas start up fund to Mohammad Haeri.

CRediT authorship contribution statement

Hongyan Gu: Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Chunxu Yang:** Software. **Shino Magaki:** Resources, Methodology, Investigation. **Neda Zarrin-Khameh:** Resources, Investigation. **Nelli S. Lakis:** Resources, Investigation. **Inma Cobos:** Resources, Investigation. **Negar Khanlou:** Resources, Investigation. **Xinhai R. Zhang:** Resources, Investigation. **Jasmeet Assi:** Resources. **Joshua T. Byers:** Resources. **Ameer Hamza:** Resources, Investigation. **Karam Han:** Resources. **Anders Meyer:** Resources. **Hilda Mirbaha:** Resources. **Carrie A. Mohila:** Resources. **Todd M. Stevens:** Resources. **Sara L. Stone:** Resources. **Wenzhong Yan:** Writing – original draft, Investigation. **Mohammad Haeri:** Writing – review & editing, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Data curation, Conceptualization. **Xiang 'Anthony' Chen:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Declaration of Generative AI in Scientific Writing

During the preparation of this work the author(s) used ChatGPT in order to check grammar mistakes and typos. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.ijhcs.2024.103315.

References

Amgad, M., Atteya, L.A., Hussein, H., Mohammed, K.H., Hafiz, E., Elsebaie, M.A., Alhusseiny, A.M., AlMoslemany, M.A., Elmatboly, A.M., Pappalardo, P.A., et al., 2022. NuCLS: A scalable crowdsourcing approach and dataset for nucleus classification and segmentation in breast cancer. GigaScience 11, giac037. http://dx.doi.org/10.1093/gigascience/giac037.

Aubreville, M., Bertram, C.A., Donovan, T.A., Marzahl, C., Maier, A., Klopfleisch, R., 2020. A completely annotated whole slide image dataset of canine breast cancer to aid human breast cancer research. Sci. Data 7 (1), 417. http://dx.doi.org/10.1038/s41597-020-00756-z.

Aubreville, M., Bertram, C., Veta, M., Klopfleisch, R., Stathonikos, N., Breininger, K., ter Hoeve, N., Ciompi, F., Maier, A., 2021. Quantifying the scanner-induced domain gap in mitosis detection. arXiv preprint arXiv:2103.16515.

Ayhan, M.S., Berens, P., 2018. Test-time data augmentation for estimation of heteroscedastic aleatoric uncertainty in deep neural networks. In: Medical Imaging with Deep Learning. https://openreview.net/forum?id=rJZz-knjz.

Ba, W., Wang, S., Shang, M., Zhang, Z., Wu, H., Yu, C., Xing, R., Wang, W., Wang, L., Liu, C., et al., 2022. Assessment of deep learning assistance for the pathological diagnosis of gastric cancer. Mod. Pathol. 35 (9), 1262–1268. http://dx.doi.org/10.1038/s41379-022-01073-z.

Balkenhol, M.C., Tellez, D., Vreuls, W., Clahsen, P.C., Pinckaers, H., Ciompi, F., Bult, P., van der Laak, J.A., 2019. Deep learning assisted mitotic counting for breast cancer. Lab. Invest. 99 (11), 1596–1606. http://dx.doi.org/10.1038/s41374-019-0275-0.

Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., Ribeiro, M.T., Weld, D., 2021. Does the whole exceed its parts? The effect of AI explanations on complementary team performance. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. CHI '21, Association for Computing Machinery, New York, NY, USA, http://dx.doi.org/10.1145/3411764.3445717.

Bertram, C.A., Aubreville, M., Gurtner, C., Bartel, A., Corner, S.M., Dettwiler, M., Kershaw, O., Noland, E.L., Schmidt, A., Sledge, D.G., Smedley, R.C., Thaiwong, T., Kiupel, M., Maier, A., Klopfleisch, R., 2020. Computerized calculation of mitotic count distribution in canine cutaneous mast cell tumor sections: Mitotic Count Is Area dependent. Vet. Pathol. 57 (2), 214–226. http://dx.doi.org/10.1177/0300985819890686, arXiv:https://doi.org/10.1177/0300985819890686. PMID: 31808382.

Bertram, C.A., Aubreville, M., Marzahl, C., Maier, A., Klopfleisch, R., 2019. A large-scale dataset for mitotic figure assessment on whole slide images of canine cutaneous mast cell tumor. Sci. Data 6 (1), 274. http://dx.doi.org/10.1038/s41597-019-0290-4, URL: https://www.nature.com/articles/s41597-019-0290-4.

Bi, W.L., Hosny, A., Schabath, M.B., Giger, M.L., Birkbak, N.J., Mehrtash, A., Allison, T., Arnaout, O., Abbosh, C., Dunn, I.F., et al., 2019. Artificial intelligence in cancer imaging: clinical challenges and applications. CA: Cancer J. Clin. 69 (2), 127–157. http://dx.doi.org/10.3322/caac.21552.

Black, N., Murphy, M., Lamping, D., McKee, M., Sanderson, C., Askham, J., Marteau, T., 1999. Consensus development methods: a review of best practice in creating clinical guidelines. J. Health Serv. Res. Policy 4 (4), 236–248. http://dx.doi.org/10.1177/135581969900400410.

Buçinca, Z., Malaya, M.B., Gajos, K.Z., 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), http://dx.doi.org/10.1145/3449287.

Bussone, A., Stumpf, S., O'Sullivan, D., 2015. The role of explanations on trust and reliance in clinical decision support systems. In: 2015 International Conference on Healthcare Informatics. pp. 160–169. http://dx.doi.org/10.1109/ICHI.2015.26.

Cai, C.J., Reif, E., Hegde, N., Hipp, J., Kim, B., Smilkov, D., Wattenberg, M., Viegas, F., Corrado, G.S., Stumpe, M.C., Terry, M., 2019a. 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. CHI '19, Association for Computing Machinery, New York, NY, USA, pp. 1–14. http://dx.doi.org/10.1145/3290605.3300234.

Cai, C.J., Winter, S., Steiner, D., Wilcox, L., Terry, M., 2019b. "Hello AI": Uncovering the onboarding needs of medical practitioners for human-AI collaborative decision-making. Proc. ACM Hum.-Comput. Interact. 3 (CSCW), http://dx.doi.org/10.1145/3359206.

Cao, S., Huang, C.-M., 2022. Understanding user reliance on AI in assisted decision-making. Proc. ACM Hum.-Comput. Interact. 6 (CSCW2), http://dx.doi.org/10.1145/3555572.

Chattopadhay, A., Sarkar, A., Howlader, P., Balasubramanian, V.N., 2018. Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks. In: 2018 IEEE Winter Conference on Applications of Computer Vision. WACV, pp. 839–847. http://dx.doi.org/10.1109/WACV.2018.00097.

Chauhan, C., Gullapalli, R.R., 2021. Ethics of AI in pathology: Current paradigms and emerging issues. Am. J. Pathol. 191 (10), 1673–1683. http://dx.doi.org/10.1016/j.ajpath.2021.06.011.

Collan, Y., Kuopio, T., Baak, J., Becker, R., Bogomoletz, W., Deverell, M., Van Diest, P., Van Galen, C., Gilchrist, K., Javed, A., Kosma, V.-M., Kujari, H., Luzi, P., Mariuzzi, G., Matze, E., Montironi, R., Scarpelli, M., Sierra, D., Sisti, S., Toikkanen, S., Tosi, P., Whimster, W., Wisse, E., 1996. Standardized mitotic counts in breast cancer evaluation of the method. Pathol. - Res. Pract. 192 (9), 931–941. http://dx.doi.org/10.1016/S0344-0338(96)80075-6, URL: https://linkinghub.elsevier.com/retrieve/pii/S0344033896800756.

Cree, I.A., Tan, P.H., Travis, W.D., Wesseling, P., Yagi, Y., White, V.A., Lokuhetty, D., Scolyer, R.A., 2021. Counting mitoses: SI(ze) matters!. Mod. Pathol. 34 (9), 1651–1657. http://dx.doi.org/10.1038/s41379-021-00825-7.

Daniel, F., Kucherbaev, P., Cappiello, C., Benatallah, B., Allahbakhsh, M., 2018. Quality control in crowdsourcing: A survey of quality attributes, assessment techniques, and assurance actions. ACM Comput. Surv. 51 (1), http://dx.doi.org/10.1145/3148148.

Del Ser, J., Barredo-Arrieta, A., Díaz-Rodríguez, N., Herrera, F., Saranti, A., Holzinger, A., 2024. On generating trustworthy counterfactual explanations. Inf. Sci.: Int. J. 655 (C), http://dx.doi.org/10.1016/j.ins.2023.119898.

Delong, C., Preuss, C.V., 2019. Black box warning.

Duregon, E., Cassenti, A., Pittaro, A., Ventura, L., Senetta, R., Rudà, R., Cassoni, P., 2015. Better see to better agree: phosphohistone H3 increases interobserver agreement in mitotic count for meningioma grading and imposes new specific thresholds. Neuro-Oncol. 17 (5), 663–669. http://dx.doi.org/10.1093/neuonc/nov002.

Efendić, E., Van de Calseyde, P.P., Evans, A.M., 2020. Slow response times undermine trust in algorithmic (but not human) predictions. Organ. Behav. Hum. Decis. Process. 157, 103–114. http://dx.doi.org/10.1016/j.obhdp.2020.01.008.

Evans, T., Retzlaff, C.O., Geißler, C., Kargl, M., Plass, M., Müller, H., Kiehl, T.-R., Zerbe, N., Holzinger, A., 2022. The explainability paradox: Challenges for xAI in digital pathology. Future Gener. Comput. Syst. 133, 281–296. http://dx.doi.org/10.1016/j.future.2022.03.009, URL: https://www.sciencedirect.com/science/article/pii/S0167739X22000838.

Ferguson, J.H., 1996. The NIH consensus development program: the evolution of guidelines. Int. J. Technol. Assess. Health Care 12 (3), 460–474. http://dx.doi.org/10.1017/S0266462300009818.

Fogliato, R., Chappidi, S., Lungren, M., Fisher, P., Wilson, D., Fitzke, M., Parkinson, M., Horvitz, E., Inkpen, K., Nushi, B., 2022. Who goes first? Influences of human-AI workflow on decision making in clinical imaging. In: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency. FAccT '22, Association for Computing Machinery, New York, NY, USA, pp. 1362–1374. http://dx.doi.org/10.1145/3531146.3533193.

Fukushima, S., Terasaki, M., Sakata, K., Miyagi, N., Kato, S., Sugita, Y., Shigemori, M., 2009. Sensitivity and usefulness of anti-phosphohistone-H3 antibody immunostaining for counting mitotic figures in meningioma cases. Brain Tumor Pathol. 26, 51–57. http://dx.doi.org/10.1007/s10014-009-0249-9.

Gal, Y., Ghahramani, Z., 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: International Conference on Machine Learning. PMLR, pp. 1050–1059.

Gaube, S., Suresh, H., Raue, M., Merritt, A., Berkowitz, S.J., Lermer, E., Coughlin, J.F., Guttag, J.V., Colak, E., Ghassemi, M., 2021. Do as AI say: susceptibility in deployment of clinical decision-aids. NPJ Digit. Med. 4 (1), 31. http://dx.doi.org/10.1038/s41746-021-00385-9.

Genzen, J.R., 2013. An overview of United States physician training, certification, and career pathways in clinical pathology (laboratory medicine). Electron. J. Int. Fed. Clin. Chem. 24 (1), 21.

Goldbrunner, R., Stavrinou, P., Jenkinson, M.D., Sahm, F., Mawrin, C., Weber, D.C., Preusser, M., Minniti, G., Lund-Johansen, M., Lefranc, F., Houdart, E., Sallabanda, K., Le Rhun, E., Nieuwenhuizen, D., Tabatabai, G., Soffietti, R., Weller, M., 2021. EANO guideline on the diagnosis and management of meningiomas. Neuro-Oncol. 23 (11), 1821–1834. http://dx.doi.org/10.1093/neuonc/noab150, arXiv:https://academic.oup.com/neuro-oncology/article-pdf/23/11/1821/41063953/noab150.pdf.

Grigorescu, S., Trasnea, B., Cocias, T., Macesanu, G., 2020. A survey of deep learning techniques for autonomous driving. J. Field Robotics 37 (3), 362–386. http://dx.doi.org/10.1002/rob.21918.

Gu, H., Haeri, M., Ni, S., Williams, C.K., Zarrin-Khameh, N., Magaki, S., Chen, X.A., 2023a. Detecting mitoses with a convolutional neural network for MIDOG 2022 challenge. In: Sheng, B., Aubreville, M. (Eds.), Mitosis Domain Generalization and Diabetic Retinopathy Analysis. Springer Nature Switzerland, Cham, pp. 211–216.

Gu, H., Huang, J., Hung, L., Chen, X.A., 2021. Lessons learned from designing an AI-enabled diagnosis tool for pathologists. Proc. ACM Hum.-Comput. Interact. 5 (CSCW1), http://dx.doi.org/10.1145/3449084.

Gu, H., Liang, Y., Xu, Y., Williams, C.K., Magaki, S., Khanlou, N., Vinters, H., Chen, Z., Ni, S., Yang, C., Yan, W., Zhang, X.R., Li, Y., Haeri, M., Chen, X.A., 2023b. Improving workflow integration with xpath: Design and evaluation of a human-AI diagnosis system in pathology. ACM Trans. Comput.-Hum. Interact. 30 (2), http://dx.doi.org/10.1145/3577011.

Gu, H., Yang, C., Al-kharouf, I., Magaki, S., Lakis, N., Williams, C.K., Alrosan, S.M., Onstott, E.K., Yan, W., Khanlou, N., Cobos, I., Zhang, X.R., Zarrin-Khameh, N., Vinters, H.V., Chen, X.A., Haeri, M., 2024. Enhancing mitosis quantification and detection in meningiomas with computational digital pathology. Acta Neuropathol. Commun. 12 (1), 7. http://dx.doi.org/10.1186/s40478-023-01707-6.

Gu, H., Yang, C., Haeri, M., Wang, J., Tang, S., Yan, W., He, S., Williams, C.K., Magaki, S., Chen, X.A., 2023c. 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, Association for Computing Machinery, New York, NY, USA, http://dx.doi.org/10.1145/3544548.3580694.

Hekler, A., Utikal, J.S., Enk, A.H., Berking, C., Klode, J., Schadendorf, D., Jansen, P., Franklin, C., Holland-Letz, T., Krahl, D., von Kalle, C., Fröhling, S., Brinker, T.J., 2019. Pathologist-level classification of histopathological melanoma images with deep neural networks. Eur. J. Cancer 115, 79–83. http://dx.doi.org/10.1016/j.ejca.2019.04.021, URL: https://www.sciencedirect.com/science/article/pii/S0959804919302758.

Holzinger, A., Carrington, A., Müller, H., 2020. Measuring the quality of explanations: the system causability scale (SCS) comparing human and machine explanations. KI-Künstliche Intell. 34 (2), 193–198. http://dx.doi.org/10.1007/s13218-020-00636-z.

Holzinger, A., Langs, G., Denk, H., Zatloukal, K., Müller, H., 2019. Causability and explainability of artificial intelligence in medicine. Wiley Interdiscip. Rev.: Data Min. Knowl. Discov. 9 (4), e1312. http://dx.doi.org/10.1002/widm.1312.

Jacobs, M., He, J., F. Pradier, M., Lam, B., Ahn, A.C., McCoy, T.H., Perlis, R.H., Doshi-Velez, F., Gajos, K.Z., 2021a. Designing AI for trust and collaboration in time-constrained medical decisions: A sociotechnical lens. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. CHI '21, Association for Computing Machinery, New York, NY, USA, http://dx.doi.org/10.1145/3411764.3445385.

Jacobs, M., Pradier, M.F., McCoy, T.H., Perlis, R.H., Doshi-Velez, F., Gajos, K.Z., 2021b. How machine-learning recommendations influence clinician treatment selections: the example of antidepressant selection. Transl. Psychiatry 11 (1), 108. http://dx.doi.org/10.1038/s41398-021-01224-x.

Jordan, M.I., Mitchell, T.M., 2015. Machine learning: Trends, perspectives, and prospects. Science 349 (6245), 255–260. http://dx.doi.org/10.1126/science.aaa8415, arXiv:https://www.science.org/doi/pdf/10.1126/science.aaa8415. URL: https://www.science.org/doi/abs/10.1126/science.aaa8415.

Kaur, H., Nori, H., Jenkins, S., Caruana, R., Wallach, H., Wortman Vaughan, J., 2020. Interpreting interpretability: Understanding data scientists' use of interpretability tools for machine learning. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. CHI '20, Association for Computing Machinery, New York, NY, USA, pp. 1–14. http://dx.doi.org/10.1145/3313831.3376219.

Van der Laak, J., Litjens, G., Ciompi, F., 2021. Deep learning in histopathology: the path to the clinic. Nat. Med. 27 (5), 775–784.

Lai, V., Liu, H., Tan, C., 2020. "Why is 'chicago' deceptive?" towards building model-driven tutorials for humans. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. CHI '20, Association for Computing Machinery, New York, NY, USA, pp. 1–13. http://dx.doi.org/10.1145/3313831.3376873.

Lai, V., Tan, C., 2019. On human predictions with explanations and predictions of machine learning models: A case study on deception detection. In: Proceedings of the Conference on Fairness, Accountability, and Transparency. In: FAT* '19, Association for Computing Machinery, New York, NY, USA, pp. 29–38. http://dx.doi.org/10.1145/3287560.3287590.

Lebedeva, A., Kornowicz, J., Lammert, O., Papenkordt, J., 2023. The role of response time for algorithm aversion in fast and slow thinking tasks. In: Artificial Intelligence in HCI: 4th International Conference, AI-HCI 2023, Held as Part of the 25th HCI International Conference, HCII 2023, Copenhagen, Denmark, July 23–28, 2023, Proceedings, Part I. Springer-Verlag, Berlin, Heidelberg, pp. 131–149. http://dx.doi.org/10.1007/978-3-031-35891-3_9.

Leichtmann, B., Humer, C., Hinterreiter, A., Streit, M., Mara, M., 2023. Effects of explainable artificial intelligence on trust and human behavior in a high-risk decision task. Comput. Hum. Behav. 139, 107539. http://dx.doi.org/10.1016/j.chb.2022.107539.

Licklider, J.C., 1960. Man-computer symbiosis. IRE Trans. Hum. Factors Electron. (1), 4–11. http://dx.doi.org/10.1109/THFE2.1960.4503259.

Lindvall, M., Lundström, C., Löwgren, J., 2021. Rapid assisted visual search: Supporting digital pathologists with imperfect AI. In: 26th International Conference on Intelligent User Interfaces. IUI '21, Association for Computing Machinery, New York, NY, USA, pp. 504–513. http://dx.doi.org/10.1145/3397481.3450681.

Litjens, G., Sánchez, C.I., Timofeeva, N., Hermsen, M., Nagtegaal, I., Kovacs, I., Hulsbergen-Van De Kaa, C., Bult, P., Van Ginneken, B., Van Der Laak, J., 2016. Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. Sci. Rep. 6 (1), 26286. http://dx.doi.org/10.1038/srep26286.

Long, D., Magerko, B., 2020. What is AI literacy? Competencies and design considerations. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. CHI '20, Association for Computing Machinery, New York, NY, USA, pp. 1–16. http://dx.doi.org/10.1145/3313831.3376727.

Louis, D.N., Perry, A., Wesseling, P., Brat, D.J., Cree, I.A., Figarella-Branger, D., Hawkins, C., Ng, H.K., Pfister, S.M., Reifenberger, G., Soffietti, R., von Deimling, A., Ellison, D.W., 2021. The 2021 WHO classification of tumors of the central nervous system: a summary. Neuro-Oncol. 23 (8), 1231–1251. http://dx.doi.org/10.1093/neuonc/noab106.

McMillan, S.S., King, M., Tully, M.P., 2016. How to use the nominal group and Delphi techniques. Int. J. Clin. Pharm. 38, 655–662.

Meyer, J.S., Alvarez, C., Milikowski, C., Olson, N., Russo, I., Russo, J., Glass, A., Zehnbauer, B.A., Lister, K., Parwaresch, R., 2005. Breast carcinoma malignancy grading by Bloom–Richardson system vs proliferation index: reproducibility of grade and advantages of proliferation index. Mod. Pathol. 18 (8), 1067–1078. http://dx.doi.org/10.1038/modpathol.3800388, URL: https://linkinghub.elsevier.com/retrieve/pii/S0893395222045781.

Min, B., Ross, H., Sulem, E., Veyseh, A.P.B., Nguyen, T.H., Sainz, O., Agirre, E., Heintz, I., Roth, D., 2023. Recent advances in natural language processing via large pre-trained language models: A survey. ACM Comput. Surv. 56 (2), http://dx.doi.org/10.1145/3605943.

Montezuma, D., Oliveira, S.P., Neto, P.C., Oliveira, D., Monteiro, A., Cardoso, J.S., Macedo-Pinto, I., 2023. Annotating for artificial intelligence applications in digital pathology: A practical guide for pathologists and researchers. Mod. Pathol. 36 (4), 100086. http://dx.doi.org/10.1016/j.modpat.2022.100086, URL: https://www.sciencedirect.com/science/article/pii/S0893395222055260.

Morrison, K., Shin, D., Holstein, K., Perer, A., 2023. Evaluating the impact of human explanation strategies on human-AI visual decision-making. Proc. ACM Hum.-Comput. Interact. 7 (CSCW1), http://dx.doi.org/10.1145/3579481.

Murphy, M., Black, N., Lamping, D., McKee, C., Sanderson, C., Askham, J., Marteau, T., 1998. Consensus development methods, and their use in clinical guideline development. Health Technol. Assess. (Winch., Engl.) 2 (3), i–88.

Nourani, M., Roy, C., Block, J.E., Honeycutt, D.R., Rahman, T., Ragan, E., Gogate, V., 2021. Anchoring bias affects mental model formation and user reliance in explainable AI systems. In: 26th International Conference on Intelligent User Interfaces. IUI '21, Association for Computing Machinery, New York, NY, USA, pp. 340–350. http://dx.doi.org/10.1145/3397481.3450639.

Pantanowitz, L., Quiroga-Garza, G.M., Bien, L., Heled, R., Laifenfeld, D., Linhart, C., Sandbank, J., Shach, A.A., Shalev, V., Vecsler, M., et al., 2020. An artificial intelligence algorithm for prostate cancer diagnosis in whole slide images of core needle biopsies: a blinded clinical validation and deployment study. Lancet Digit. Health 2 (8), e407–e416. http://dx.doi.org/10.1016/S2589-7500(20)30159-X.

Park, J.S., Barber, R., Kirlik, A., Karahalios, K., 2019. A slow algorithm improves users' assessments of the algorithm's accuracy. Proc. ACM Hum.-Comput. Interact. 3 (CSCW), http://dx.doi.org/10.1145/3359204.

Passi, S., Vorvoreanu, M., 2022. Overreliance on AI literature review. Microsoft Res..

Pena, G.P., Andrade-Filho, J.S., 2009. How does a pathologist make a diagnosis? Arch. Pathol. Lab. Med. 133 (1), 124–132. http://dx.doi.org/10.5858/133.1.124.

Plass, M., Kargl, M., Kiehl, T.-R., Regitnig, P., Geißler, C., Evans, T., Zerbe, N., Carvalho, R., Holzinger, A., Müller, H., 2023. Explainability and causability in digital pathology. J. Pathol.: Clin. Res. 9 (4), 251–260. http://dx.doi.org/10.1002/cjp2.322, URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/cjp2.322. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/cjp2.322.

Pohn, B., Kargl, M., Reihs, R., Holzinger, A., Zatloukal, K., Müller, H., 2019. Towards a Deeper Understanding of How a Pathologist Makes a Diagnosis: Visualization of the Diagnostic Process in Histopathology. In: 2019 IEEE Symposium on Computers and Communications. ISCC, pp. 1081–1086. http://dx.doi.org/10.1109/ISCC47284.2019.8969598, URL: https://ieeexplore.ieee.org/document/8969598. ISSN: 2642-7389.

Rastogi, C., Zhang, Y., Wei, D., Varshney, K.R., Dhurandhar, A., Tomsett, R., 2022. Deciding fast and slow: The role of cognitive biases in AI-assisted decision-making. Proc. ACM Hum.-Comput. Interact. 6 (CSCW1), http://dx.doi.org/10.1145/3512930.

Regitnig, P., Müller, H., Holzinger, A., 2020. Expectations of Artificial Intelligence for Pathology. In: Holzinger, A., Goebel, R., Mengel, M., Müller, H. (Eds.), Artificial Intelligence and Machine Learning for Digital Pathology: State-of-the-Art and Future Challenges. Springer International Publishing, Cham, pp. 1–15. http://dx.doi.org/10.1007/978-3-030-50402-1_1.

Schemmer, M., Hemmer, P., Nitsche, M., Kühl, N., Vössing, M., 2022. A meta-analysis of the utility of explainable artificial intelligence in human-AI decision-making. In: Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society. AIES '22, Association for Computing Machinery, New York, NY, USA, pp. 617–626. http://dx.doi.org/10.1145/3514094.3534128.

Schemmer, M., Kuehl, N., Benz, C., Bartos, A., Satzger, G., 2023. Appropriate reliance on AI advice: Conceptualization and the effect of explanations. In: Proceedings of the 28th International Conference on Intelligent User Interfaces. IUI '23, Association for Computing Machinery, New York, NY, USA, pp. 410–422. http://dx.doi.org/10.1145/3581641.3584066.

Stacke, K., Eilertsen, G., Unger, J., Lundström, C., 2020. Measuring domain shift for deep learning in histopathology. IEEE J. Biomed. Health Inform. 25 (2), 325–336. http://dx.doi.org/10.1109/JBHI.2020.3032060.

Surden, H., 2019. Artificial intelligence and law: An overview. Georgia State Univ. Law Rev. 35, 19–22.

Tan, M., Le, Q., 2019. EfficientNet: Rethinking model scaling for convolutional neural networks. In: Chaudhuri, K., Salakhutdinov, R. (Eds.), Proceedings of the 36th International Conference on Machine Learning. In: Proceedings of Machine Learning Research, vol. 97, PMLR, pp. 6105–6114, URL: https://proceedings.mlr.press/v97/tan19a.html.

Taze, D., Hartley, C., Morgan, A.W., Chakrabarty, A., Mackie, S.L., Griffin, K.J., 2022. Developing consensus in histopathology: the role of the delphi method. Histopathology 81 (2), 159–167. http://dx.doi.org/10.1111/his.14650.

Van Bergeijk, S.A., Stathonikos, N., Ter Hoeve, N.D., Lafarge, M.W., Nguyen, T.Q., Van Diest, P.J., Veta, M., 2023. Deep learning supported mitoses counting on whole slide images: A pilot study for validating breast cancer grading in the clinical workflow. J. Pathol. Inform. 14, 100316. http://dx.doi.org/10.1016/j.jpi.2023.100316, URL: https://linkinghub.elsevier.com/retrieve/pii/S215353923001 30X.

Van de Ven, A.H., Delbecq, A.L., 1972. The nominal group as a research instrument for exploratory health studies. Am. J. Public Health 62 (3), 337–342. http://dx.doi.org/10.2105/ajph.62.3.337.

Vasconcelos, H., Jörke, M., Grunde-McLaughlin, M., Gerstenberg, T., Bernstein, M.S., Krishna, R., 2023. Explanations can reduce overreliance on AI systems during decision-making. Proc. ACM Hum.-Comput. Interact. 7 (CSCW1), http://dx.doi.org/10.1145/3579605.

Veale, M., Zuiderveen Borgesius, F., 2021. Demystifying the draft EU artificial intelligence act—Analysing the good, the bad, and the unclear elements of the proposed approach. Comput. Law Rev. Int. 22 (4), 97–112.

Veta, M., Van Diest, P.J., Jiwa, M., Al-Janabi, S., Pluim, J.P.W., 2016. Mitosis Counting in Breast Cancer: Object-Level Interobserver Agreement and Comparison to an Automatic Method. PLoS ONE 11 (8), e0161286. http://dx.doi.org/10.1371/journal.pone.0161286, URL: https://dx.plos.org/10.1371/journal.pone.0161286.

Wang, D., Khosla, A., Gargeya, R., Irshad, H., Beck, A.H., 2016. Deep learning for identifying metastatic breast cancer. arXiv preprint arXiv:1606.05718.

Wang, W., Zhao, Y., Teng, L., Yan, J., Guo, Y., Qiu, Y., Ji, Y., Yu, B., Pei, D., Duan, W., Wang, M., Wang, L., Duan, J., Sun, Q., Wang, S., Duan, H., Sun, C., Guo, Y., Luo, L., Guo, Z., Guan, F., Wang, Z., Xing, A., Liu, Z., Zhang, H., Cui, L., Zhang, L., Jiang, G., Yan, D., Liu, X., Zheng, H., Liang, D., Li, W., Li, Z.-C., Zhang, Z., 2023. Neuropathologist-level integrated classification of adult-type diffuse gliomas using deep learning from whole-slide pathological images. Nature Commun. 14 (1), 6359. http://dx.doi.org/10.1038/s41467-023-41195-9.

Wolfe, J.M., Horowitz, T.S., Van Wert, M.J., Kenner, N.M., Place, S.S., Kibbi, N., 2007. Low target prevalence is a stubborn source of errors in visual search tasks. J. Exp. Psychol.: Gen. 136 (4), 623.

Zhang, Z., Chen, P., McGough, M., Xing, F., Wang, C., Bui, M., Xie, Y., Sapkota, M., Cui, L., Dhillon, J., Ahmad, N., Khalil, F.K., Dickinson, S.I., Shi, X., Liu, F., Su, H., Cai, J., Yang, L., 2019. Pathologist-level interpretable whole-slide cancer diagnosis with deep learning. Nat. Mach. Intell. 1 (5), 236–245. http://dx.doi.org/10.1038/s42256-019-0052-1.

Zhang, Y., Liao, Q.V., Bellamy, R.K.E., 2020. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. In: FAT* '20, Association for Computing Machinery, New York, NY, USA, pp. 295–305. http://dx.doi.org/10.1145/3351095.3372852.

Zhou, S., Pfeiffer, N., Islam, U.J., Banerjee, I., Patel, B.K., Iquebal, A.S., 2022. Generating counterfactual explanations for causal inference in breast cancer treatment response. In: 2022 IEEE 18th International Conference on Automation Science and Engineering. CASE, pp. 955–960. http://dx.doi.org/10.1109/CASE49997.2022.9926519.