

# Rapid Assisted Visual Search

## Supporting Digital Pathologists with Imperfect AI

Martin Lindvall\*  
martin@ixd.ai  
Linköping University  
Sweden

Claes Lundström\*  
claes.lundstrom@liu.se  
Linköping University  
Sweden

Jonas Löwgren  
jonas.lowgren@liu.se  
Linköping University  
Sweden

### ABSTRACT

Designing useful human-AI interaction for clinical workflows remains challenging despite the impressive performance of recent AI models. One specific difficulty is a lack of successful examples demonstrating how to achieve safe and efficient workflows while mitigating AI imperfections. In this paper, we present an interactive AI-powered visual search tool that supports pathologists in cancer assessments. Our evaluation with six pathologists demonstrates that it can 1) reduce time needed with maintained quality, 2) build user trust progressively, and 3) learn and improve from use. We describe our iterative design process, model development, and key features. Through interviews, design choices are related to the overall user experience. Implications for future human-AI interaction design are discussed with respect to trust, explanations, learning from use, and collaboration strategies.

### CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in interaction design**; • **Applied computing** → **Life and medical sciences**; • **Computing methodologies** → *Machine learning*.

### KEYWORDS

machine learning, human-AI interaction design, explainable AI

#### ACM Reference Format:

Martin Lindvall, Claes Lundström, and Jonas Löwgren. 2021. Rapid Assisted Visual Search: Supporting Digital Pathologists with Imperfect AI. In *26th International Conference on Intelligent User Interfaces (IUI '21)*, April 14–17, 2021, College Station, TX, USA. ACM, New York, NY, USA, 10 pages. <https://doi.org/10.1145/3397481.3450681>

## 1 INTRODUCTION

Recent machine learning (ML) techniques have achieved near expert-level predictive accuracy on several medical imaging tasks [1, 21, 34]. However, certain shortcomings make current probabilistic AI models imperfect and create challenges for translating experimental results into clinical workflows. For instance, unknown subsets of cases in real practice can cause unpredictable errors [24, 28]. Due to variability in manual processes and equipment changes, the

predictive capability can also decay over time [27, 32]. As an alternative to computational methods for creating models with less bias and better generalizability, successful user interfaces for human-AI interaction could help physicians deal with these “imperfections” [6] while still employing the technology to increase positive patient outcomes.

Successful human-AI interaction is, however, difficult to design due to the need to balance trade-offs between exploiting AI capabilities and mitigating its imperfections [39]. Designers report being hampered by a lack of design patterns, prototyping methods and examples of user interfaces matching AI capabilities to user value [11, 37, 38].

This paper presents Rapid Assisted Visual Search (RAVS) – a human-AI interface that utilises an imperfect model’s strengths to aid digital pathologists in assessing colorectal cancer. Specifically, we address the task of searching regional lymph nodes for signs of tumour metastasis, a process routinely performed after surgical resections [10].

To create RAVS, we adopted a design-led approach where ideas on technical properties and interactive behaviours were prototyped, refined, and assessed in an iterative process in collaboration with pathologists. The process included data collection from retrospective cases, model development, workplace observations, interaction design, and system development. We describe RAVS through its computational components and user interface features. To concisely motivate features, we annotate them with insights gained from the explorative design process.

Evaluation with six pathologists demonstrates that RAVS can reduce review time with maintained quality. Our interviews reveal that users could adapt their trust progressively by being in control of decision-making, sensitivity, and the AI-features used. Additionally, human-AI systems that allow user-corrections of predictions can theoretically improve through the data collected in use. However, care must be taken to incentivise such action without negatively affecting the user experience [19, 40]. We designed RAVS to make this aspect effortless. In a computational evaluation, we show that the usage logging in RAVS can indeed improve predictive accuracy on future cases.

In summary, the main contributions of this paper are:

- Rapid Assisted Visual Search (RAVS), a human-AI user interface for low-prevalence search tasks that is compatible with imperfect AI models and can learn from being used
- An empirical evaluation showing that our system is fast, accurate and helpful in aiding pathologists assessing colorectal lymph nodes for signs of tumour metastasis

\*Also with Sectra AB.



This work is licensed under a Creative Commons Attribution International 4.0 License.

IUI '21, April 14–17, 2021, College Station, TX, USA  
© 2021 Copyright held by the owner/author(s).  
ACM ISBN 978-1-4503-8017-1/21/04.  
<https://doi.org/10.1145/3397481.3450681>

## 2 RELATED WORK

### 2.1 Artificial Intelligence for Digital Pathology

The CAMELYON challenge and the associated datasets attracted numerous ML-researcher’s to the task of identifying tumour metastasis in lymph nodes [2, 20]. Most high-scoring approaches utilise machine learning using convolutional neural networks (CNN), sometimes adding random forests for the final classification [2]. One particularly strong approach has been the LYNA [21] method that overall performed well, but produced occasional failures due to out-of-focus or scanning artefacts. Similar CNN-based approaches have demonstrated impressive performance in experimental settings for other digital pathology tasks, such as, Ki67-scoring in breast carcinomas [18], performing Gleason grading for prostate biopsies [5] and identifying basal cell carcinomas [8]. However, there is relatively little evidence to assess whether and how these could be useful in a clinical setting where there might exist sociotechnical challenges [3], more considerable variations in image quality and rare target phenomena.

In some systems, the predictive components only partly aid the task, assisting the operator in what is sometimes called human-in-the-loop systems. In a study of one such system, the efficiency and quality effects of using the LYNA model with an AI-powered user interface was evaluated and shown to reduce review time, but only for a small subset of cases [33]. The assistive system consisted of a zoomable user interface visualising regions of interest using bounding boxes, color-coded for prediction certainty. In a system for Gleason grading, AI assistance increased concordance for most observers [4]. In a study of a content-based retrieval system for pathologists, the system demonstrated potential to resolve some difficult decisions by allowing users to refine results interactively [6]. While these works confirm that human-AI interaction can yield improvements in some situations, the larger factors contributing to success remain mostly unknown. Further work is needed to understand the gap between potential benefit in experimental settings and systems that can improve patient outcomes in clinical use.

### 2.2 Designing Human-AI interaction

A growing body of research addresses more general design and human-computer interaction challenges in using predictive models to interactively aid users. For instance, the presentation style of decision aids can affect human-AI ensemble performance [15]. Some have suggested guidelines that aid designers when seeking desirable properties and behaviours, including recommendations for treating prediction failures [1, 12, 30]. Within the explainable AI domain, the black-box nature of machine learning-based models is commonly identified as a threat to user trust and decision accountability, and several explanatory strategies have been proposed to mitigate this issue [7, 17, 26, 36].

Traditional user-centered design methods are challenging to apply to human-AI interactions, partly due to the predictive component being central to the overall experience and a lack of viable low-fidelity prototyping methods to simulate realistic prediction behaviour [11, 39]. Few works present high fidelity design concepts or user interface patterns that are explicitly useful over a more extensive range of human-AI interactive systems and associated tasks.

Especially relevant for our targeted search task, Forlines and Balakrishnan [14] presented and evaluated three techniques, all based on segmenting objects and recombining them in novel ways. Notably, a grid-based layout slightly improved error rates on low-prevalence tasks without a speed penalty, while a Rapid Serial Visual Presentation [31] technique further improved error rates, but at the cost of search time.

## 3 DESIGN: RAPID ASSISTED VISUAL SEARCH

In this section we describe the development of our predictive model and the design of the user interface that aims to utilise the strengths of the model together with a human in the loop for a visual search task.

### 3.1 Background: Assessing Lymph Nodes for Tumour Metastasis

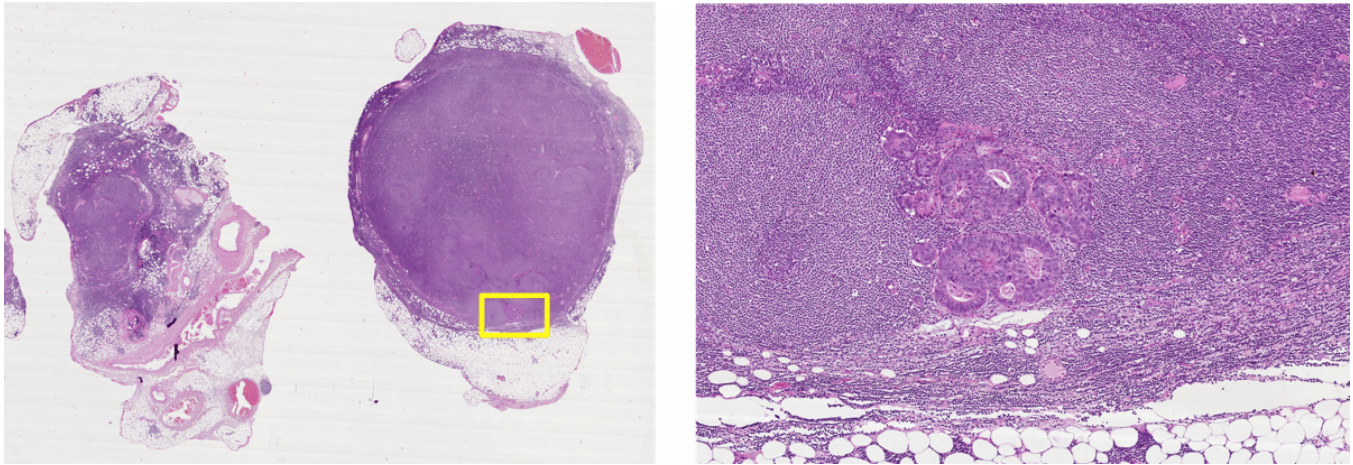
Colorectal cancer is the third most common cancer worldwide and second in terms of mortality [9]. The most common treatment is surgical removal of the tumour and surrounding tissue. During this procedure, regional lymph nodes are also removed to assess whether the tumour has spread to the lymphatic system. If the tumour has spread, the patient will be offered chemotherapy or other adjuvant therapies to increase life expectancy.

As part of regular procedure, at least 12 lymph nodes are extracted, cut into multiple thin sections, placed on glass slides and subjected to microscopic examination by a pathologist [10]. In recent years, several clinics have started using high-resolution scanners and viewing the resulting gigapixel images through a zoomable user interface on a computer workstation instead of a microscope [35]. For an example, see Figure 1.

Assessing lymph nodes for signs of tumour metastasis can be characterised as a search task where the goal is, for each lymph node, to determine whether a target (a tumour-positive region) is either present or absent. If a target is found, the search can immediately continue to the next lymph node. Positive regions can be small. Thus, to determine target-absent, the pathologist will need to systematically review the entire specimen in high magnification or use some internal threshold of “done” to decide when they have reviewed it in enough detail. When all nodes are examined, the lymph node status is reported as the number of nodes examined and number of nodes positive.

In general, search tasks can be characterised by the frequency of targets. In the clinically representative dataset used for this study, only five percent of sections were positive. Low prevalence, when the sought object is rare, has been associated with a detrimental effect on sensitivity [13].

To perform this assessment of lymph nodes in a typical digital pathology system, pathologists must click each image of a case, identify and navigate to a lymph node section and then search the zoomable image for signs of cancer. During this process the pathologist must keep track of the number of normal versus positive lymph nodes observed. The process is visually depicted in the top part of Figure 3. Determining lymph node status is only one part of the work that the pathologists perform on the case, which also includes grading and classifying the primary tumour and documenting any accidental findings.



**Figure 1: Scanned lymph nodes. (left) Two sections cut from resected lymph nodes are shown at 1X magnification. A yellow rectangle indicates the viewport shown to the right, which shows a tumour region in 10X magnification.**

### 3.2 Design Goal

Our primary goal was to support pathologists when assessing lymph nodes for signs of tumour metastasis, intending to help them classify each node as either positive (target-present) or negative (target-absent) and report the total. In this process, the quality of the outcome and the speed at which decisions can be made are important. Hence, when compared to the current manual approach, our design might improve speed, quality or both. We derived secondary desiderata from what might be barriers to clinical adoption of AI-assistance:

- That the pathologist feels in control and confident in the decision-making.
- That the pathologist can explain how they received support for a particular case.
- That the pathologist can choose the level of support by assessing when the tool should be used or not.

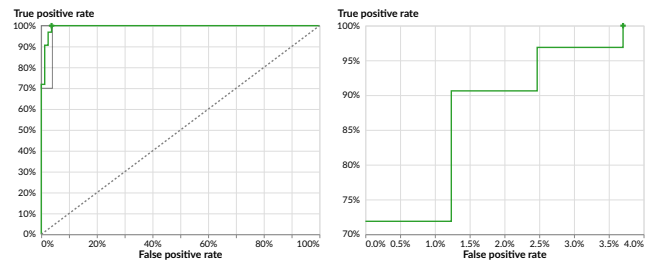
Additionally, we also sought to design the interaction so that the AI component could learn from pathologists through use.

### 3.3 Design Process

Overall a user-centered design methodology was used in combination with computational experimentation, including data extraction, annotation and supervised machine learning. During the design phase, we used both low- and high-fidelity prototyping. However, our experiences with paper-based sketches led us to use interactive prototypes earlier and to a higher degree than is typical for non-AI-based designs. One primary collaborator, a junior pathologist, provided formative input in monthly sessions for eight months. In addition, we performed formative evaluations of interim high-fidelity prototypes. One version was evaluated with three pathologists, and a later version evaluated by two pathologists. We used an explorative design approach to discover requirements and iteratively construct an interactive assistive tool.

### 3.4 Model development

The behaviour and achievable performance of predictive models can be hard to prototype using low-fidelity methods [11], and our experiences second this. To reach similar performance as reported from prior studies on breast lymph nodes, we wanted a “good enough” tumour classification model for colorectal lymph nodes. As no dataset was readily available, we extracted 977 images from 39 patients, and one pathologist annotated normal and tumour areas. The dataset, including images and annotations, has been published separately as AIDA-LNCO [22]. We applied supervised machine learning using a simple four-layer convolutional neural network with standard colour and geometrical data augmentations. As described in the next section, our user interface depends on relative rather than absolute prediction values. Thus, to avoid over-confidence, we added an entropy penalty to our loss function [25]. Additionally, we combined eight models trained with small variations in hyperparameters into an ensemble. The model achieved an area under the ROC curve (AUC) of 99.5% on a test set, see Figure 2. Operating at perfect sensitivity, that corresponds to an accuracy of 97.3%.



**Figure 2: Receiver operating characteristic (ROC) curve for our model. (left) The green line shows our model’s ability to discriminate positive from negative sections at varying discrimination thresholds. (right) Same curve focused on the top-left corner.**

Our model's performance is comparable to other works on identifying tumours in lymph nodes. For instance, LYNA achieved an AUC of 99.6% on sentinel lymph nodes from breast [21]. While this predictive accuracy can appear high, the number of false positives is likely unacceptable for clinical use. For instance, in the study by Steiner. et al. assessing an AI-assistant for breast lymph nodes, all but one pathologist had perfect specificity in the manual condition [33].

### 3.5 System Description

Rapid Assisted Visual Search (RAVS) is conceptualised as a tool that can be opened by the pathologist as part of her regular digital pathology system. We chose to situate RAVS as an opt-in tool since the assessment of lymph nodes is only one part in the larger process of reviewing a case from a colorectal resection.

In RAVS, the use of ML-models assists the user in multiple non-overlapping ways:

- Navigation between lymph node sections across the case, regardless of which image they were captured in
- Navigation to AI-suggested regions that need to be reviewed to determine tumour-present or absent (normal)
- Visualisation of progress throughout the case, including the regions or sections that will be reviewed next
- Keeping track of the current count of positive, normal and unreviewed sections

We intentionally designed the system not to recommend decisions or reveal the AI's prediction for sections and suggested regions, in order not to bias the pathologist and enhance their sense of control. The manual and assisted workflows are summarised in Figure 3.

The user interface of RAVS, depicted in Figure 4, has an overview-detail layout with lymph node sections and detected regions of interest in the overview pane. The detail pane holds a zoomable user interface, with a viewing position controlled by the overview pane. At the beginning of a new case, the tool selects the first lymph node section, which fills the entire detail pane. The primary means of navigation is to mark the viewed area as target-present or absent by pressing '1' or 'space', respectively. As soon as any part of the section is marked as positive (target-present), the user is immediately moved to the next section. For each new section, the view always starts at an overview showing the entire section. If 'space' is pressed, the user will continue to the first region of interest within the current section. If all regions of the section are marked as negative, the section is marked as negative and navigation continues to the next section in the overview. The tool shows the current tally of positive, negative and unseen sections. All navigation operations are visualised through short zoom-and-pan animations that fluidly move the viewport to the new positions. Additionally, the user can mark sections or regions as 'later', which defers decisions but makes them easy to return to at a later point.

The computational part of the system consists of lymph node segmentation that identifies sections on each image, tumour classification that results in a positivity map, region of interest segmentation that determines the extent of each positive region in the positivity map and finally, region scoring that determines which regions are shown, and in what order.

Overall, the design intent has balanced trade-offs between opportunities and challenges uncovered during the iterative design process. We next briefly describe some specific features together with the insights from the design process that motivated them. The system implements some well-known HCI principles and concepts: **Rapid Serial Visual Presentation** [14, 31], **Overview first, then details on demand** [29] and **Reversible actions** [30].

Furthermore, the system implements several human-AI interaction-specific features that are less established. We call our key features: **Manual any time**, **Visualise just enough detail**, **Hide underlying probabilities**, **Always show regions**, **Order by probability**, **Incremental sensitivity** and finally **Learn from use**.

**3.5.1 Rapid serial visual presentation.** Allow for rapid navigation using the keyboard. Place phenomena of interest in the same screen position to avoid excessive eye movement. *Motivation:* Afford a time-efficient and ergonomic workflow, without losing accuracy.

**3.5.2 Overview first, then details on demand.** Present the user interface in an overview-detail layout. In addition, during RSVP, present section overviews for each new section before continuing to detected regions. *Motivation:* Keep track of where you are. If the classifier fails, allow using section navigation and counting aid only, ignoring suggested regions.

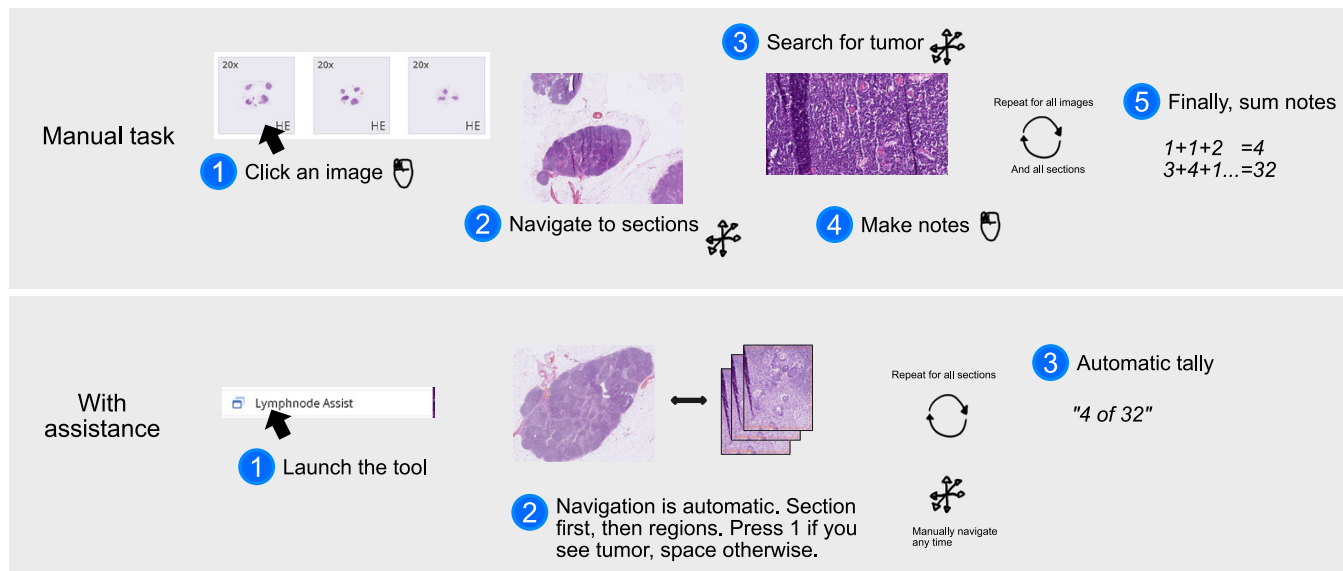
**3.5.3 Reversible actions.** Allow going back through the overview panel or by pressing the back arrow on the keyboard. *Motivation:* It is known that RSVP is associated with errors due to being too fast [13].

**3.5.4 Manual any time.** The tool is designed as an opt-in tool. The user can navigate and use other tools at any time. Pressing one of the RSVP keys continues the rapid sequence. *Motivation:* Reality is messy. The model or other components may fail unexpectedly, requiring a fallback option. The work might be interrupted. Additionally, there might not be enough trust to use RAVS.

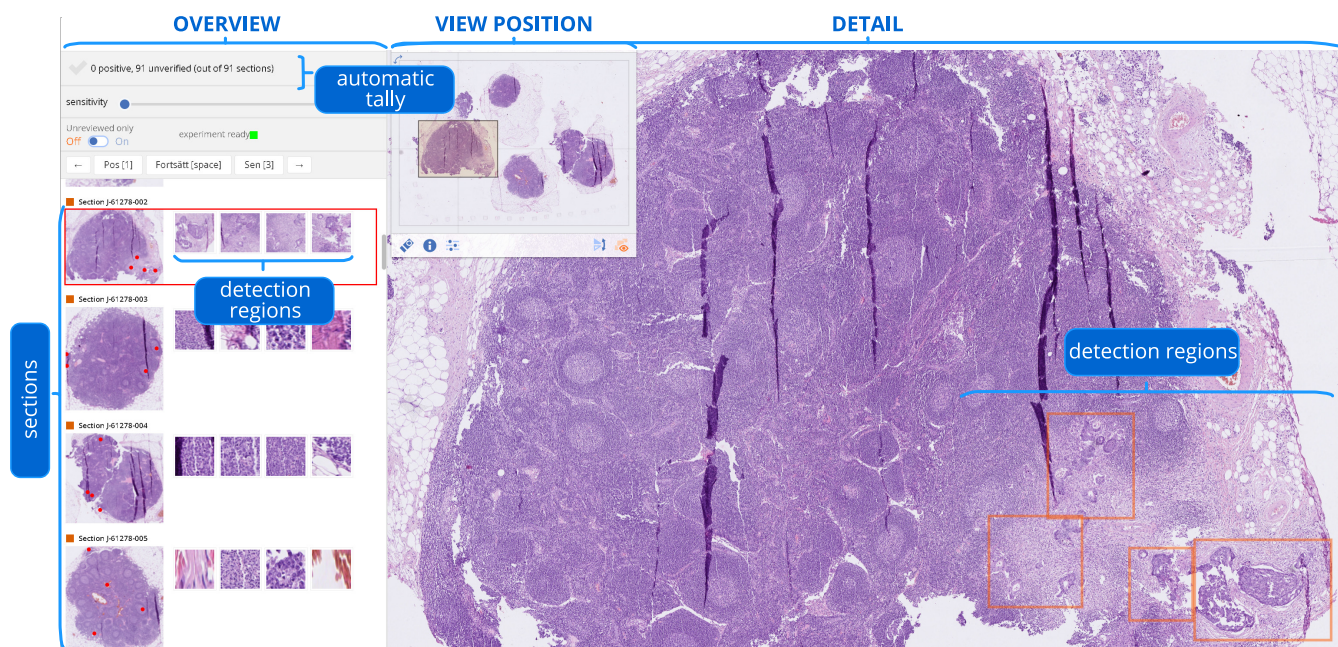
**3.5.5 Visualise just enough AI detail.** Only show low contrast boxes around a few regions of interest. Make region of interest algorithm consider both the pixel predictions and the viewport that a human can see without excessive eye movement. Draw the box around that viewport, rather than the bounding box of positive pixels. *Motivation:* Too specific detection regions, e.g., showing detections as polygons or visualising output as a heatmap, might decrease trust and drive behaviour that overemphasises details of the prediction rather than keeping the medical image as the center of attention.

**3.5.6 Hide underlying probabilities.** Do not show uncertainty or probability explicitly, e.g., through colour-coded boxes, as percentages or a recommended label. *Motivation:* Confirmation bias might lead to reduced specificity. Also, ensure that pathologists feel in control of the final decision.

**3.5.7 Always show regions.** Always show N regions even if they are likely normal. Initial N can be derived from the performance of the model depending on needed sensitivity. Show the regions with the highest probability of containing tumour, even if they are close to zero. *Motivation:* Filtering detections at some threshold would lead to many normal sections without any detections. If the trust in



**Figure 3: Comparison of manual (top) and assisted (bottom) workflows.** During manual review the pathologist clicks each image, locates lymph nodes, searches each node for cancer and keeps a tally of present and absent findings. In the assisted workflow using RAVS, the pathologist marks viewed areas as target-present or absent, navigation and tally are automatic.



**Figure 4: RAVS User interface.** An overview of sections and selected regions are shown on the left. On the right side, a detail pane shows the scanned image in a zoomable user interface. Selected regions are also shown here as bounding boxes with a low-contrast orange color. At the top left of the detail pane, the current position of the viewport is shown in relation to the current image.

the system is not enough, the user might revert to slower manual review.

**3.5.8 Order by probability.** Order detected regions within each section by the probability assigned by the model. Ensure that model training does not produce overconfident models. *Motivation:* Afford a time efficient workflow; if a positive region is confirmed, the user can proceed to the next section. Additionally, when underlying probabilities are hidden, the user needs some implicit way of assessing the quality of predictions in order to adapt their trust.

**3.5.9 Incremental sensitivity.** Allow controlling sensitivity through the fixed number of shown detection regions. A smaller N is used initially. If the case has many positive sections, stop after one round through the case. If no or few positives are found, the user can increase the sensitivity and go through negatives once more. *Motivation:* Pathologists need varying precision depending on the case at hand. If a case already has a high number of positive lymph nodes, finding more will not affect treatment decisions. If the case appears negative, increasing sensitivity is only one of several options, which also include re-reviewing suspicious regions marked as 'later' or ordering additional immunohistochemical treatment of the tissue. Also, controlling sensitivity this way allows adaptation to models of varying performance.

**3.5.10 Learn from use.** Provide an automatic tally. *Motivation:* Since the tool can be used with varying automatic support, create an incentive for marking both positives and negatives using the tool so that training data can be collected in use.

## 4 EVALUATION

### 4.1 Data

We wanted to evaluate RAVS on cases that reflect regular clinical review that a deployed system might encounter. We retrospectively extracted 50 chronologically consecutive cases from patients that had confirmed colorectal adenocarcinomas and had undergone surgical removal with subsequent review of lymph nodes. In order to reflect a dataset shift that might occur, we set the start period to be one year after the most recent case in the dataset that was used for model training. This dataset has separately been published as AIDA-LNCO2 [23]. From the set of 50 consecutive cases, we randomly sampled 14 cases for use in the user study. After the experiment, we assessed ground truth showing that out of the 14 cases, seven were positive in the sense that they had at least one positive section. There were in total 675 lymph node sections, out of which 35 were positive, corresponding to a target prevalence of 5.2%.

### 4.2 Participants

We contacted ten pathologists or pathology residents. One declined directly. Nine responded that they could participate, and six completed the study. Participants received no reimbursement. Four participants were pathology specialists. Two participants were pathology residents, one in their final year and the other in the second year. The second-year resident had over five years of microscopy experience from another specialisation. Out of the four pathology specialists, two had less than five years of experience, and two had

over ten years of experience. The participants were from four different medical centres in Sweden. Four of the participants worked in departments where they used digital pathology systems daily.

### 4.3 Method

The experiment placed participants in a balanced within-subject design reviewing all cases both manually and using RAVS in an assisted condition. The washout period was at least one week. We measured efficiency as time taken per section, derived from usage logs. To measure quality, we asked participants to count and score sections individually as positive or negative. The ground truth was compiled after all participants had finished the experiment. For each section, a full consensus was accepted as truth. Sections without consensus were sent to one senior expert in gastrointestinal pathology who determined the truth.

The manual condition was performed using a commercial clinic-grade digital pathology viewer without any AI assistance. The experiment was self-paced and done unattended over a web interface. Participants received training material in both written form and as video recordings. The material covered the experiment, the assistive tool and the underlying model, including its performance on the test set. Participants were also given five cases to train upon before commencing the study.

In the assisted condition, we did not allow manual control of the sensitivity setting, as it might make the quality-speed trade-off difficult to analyse. Instead, the cases were initially shown with four regions. In order to gain qualitative insights into the use of incremental sensitivity, when the participant had finished the case, they were shown a dialog asking whether they would like to re-review the case at a higher sensitivity. Regardless of their answer, a higher sensitivity setting showing eight regions was always enabled. Unless otherwise stated, we report efficiency and quality outcomes of the first iteration with four regions.

Additionally, we scored the model in an autonomous AI condition. We used the maximum probability to score sections. The discrimination threshold value was set to the one yielding the minimum number of false positives at a sensitivity above 97%, as observed on the test set of our model training.

After the experiment, we conducted one-hour semi-structured interviews. In the first part, questions centred on trust and patient safety, viability for clinical practice and the overall experience. In the second part, we showed a few design alternatives to spark discussions on automation issues such as confirmation bias and being out of the loop. Alternatives included a prototype where high-probability positive regions were explicitly highlighted and where very low-probability regions had been hidden altogether. We also showed probability heatmaps of normal cases and a case where the model's output was problematic. Audio recordings were transcribed and analysed inductively for salient insights on the interaction experience.

We analysed the relationship between condition and time taken through a mixed effects model in R. Time taken per section was the dependent variable, condition and session (first or second) were fixed effects. Images were treated as random intercepts, subjects-condition as random slope. P-values were obtained by the likelihood ratio test using the anova function in R, comparing the full model

to a null model with the fixed effect of condition removed. We did not perform statistical inference on quality outcomes.

Half of the time-measurements from one participant were excluded due to severe network issues during one of the four blocks. We excluded two images (four sections) from the analysis altogether since they contained tumour deposits and not lymph nodes, a scenario not covered in the instructions.

## 4.4 Findings

We report quantitative measurements in terms of quality and efficiency. Next, findings from interviews are related to RAVS key features and grouped by four identified themes: Transparency and trust, Workflow and control, Algorithms in the loop and Towards clinical use.

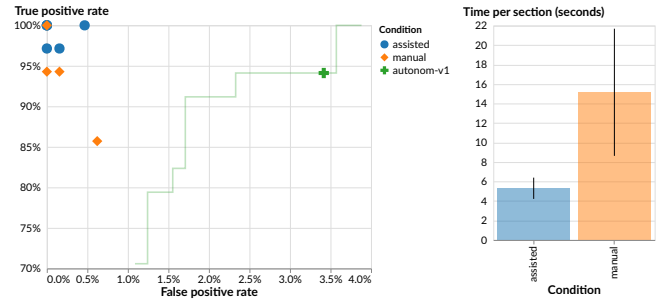
**4.4.1 Quality and efficiency.** The mean sensitivity in the manual condition was 95.2% compared to 99% in the assisted condition. Specificity was 99.9% in both conditions. In the assisted condition, two participants missed one positive section. Most positives were detectable in the first suggested region. For the 35 positive sections, 31 were identified as tumour in the first region, two in the second, one in the third and finally one in the fifth. Results are presented in Figure 5. We also evaluated the model in an autonomous condition using a pre-determined threshold. All participants outperformed the model in terms of specificity in both manual and assisted conditions. In terms of sensitivity, one junior pathologist scored below the model in the manual condition.

In terms of efficiency, using RAVS is significantly faster ( $p < 0.05$ ) than manual review, corresponding to only needing 35% of manual reviewing time. When using a sensitivity showing eight instead of four regions, the mean time was on average increased with a factor 1.75, corresponding to the second pass adding 80% of the time of the first pass.

In summary, RAVS enabled participants to utilise the AI's predictions to work faster with maintained quality, even though the predictive accuracy of the model was below pathologist levels.

**4.4.2 Transparency and trust.** By **always showing regions**, participants were by repeated exposure able to explicate patterns in regions by correlating them to the phenomena detected as well as point to some problematic positives. They also formed an opinion on the trustworthiness of the model by utilising the **order** in which regions were presented. For example, they commonly mentioned the order of positives: "True tumour is always in the first region" (P1) and explicated characteristics of normal regions: "Typical false positives are germinal centers and histiocytes. They can look a little weird. Novice pathologists usually react to these as well. It feels reasonable if it must select something in a normal that it goes with the germinal centers" (P2). Some participants formed new strategies based on these learned characteristics, exploiting them for speed: "Whenever it was showing me detections only in germinal centers in the overview, I would just step through it really quickly. Click-click-click. I wish there were an option to accept it as negative straight away." (P4)

However, they struggled with determining what would indicate a risk for an individual lymph node section and what could indicate that the model had some global flaw. They sought to understand potential risks but lacked certainty in identifying a malfunction:



**Figure 5: Quality and efficiency results.** The left part shows the true and false positive rates of six participants under manual (orange diamonds) and assisted (blue circles) conditions. The figure also includes the result of the model at a pre-defined threshold (green cross) and the effect of varying the discrimination threshold (green line). Three participants achieved a perfect score in both conditions. To the right, the average review time per section in seconds are shown for manual and assisted conditions. Error bars indicate 95% confidence intervals.

"You are sometimes surprised when it points out areas where there can be nothing, like fat, or some dirt on the glass. Then one thinks, if this was in real use, I would worry. It feels like it is looking at the wrong things." (P2). "It is picking thick slices, those could look like artifacts and a tumour could be hiding, I don't know if that's a risk or not" (P1). As participants correctly concluded, these situations could indicate that the model was failing, or it could just be a consequence of the section being fully normal. Without further information, users were on their own to decide whether to fall back to manual review or risk missing a positive.

**Incremental sensitivity**, going through the case in multiple passes with increasing sensitivity, built trust for sceptics: "I am a bit suspicious, like, how can I be sure if I don't look at everything. In the beginning I answered yes, but later I was like - no - this is good enough. I changed my mind faster than I thought I would by going through it a few times." (P5). Others used it as a precaution "I answered Yes, against my gut feeling, just to be confident" (P3). However, having the second pass forced upon them was perceived as a nuisance to those who had already placed high trust in the system: "Going through 280 regions was boring, having to do it again, then it's even more boring than manual review" (P1). Only one participant explicitly mentioned using it for the purpose we had intended - to be sure of all-negative cases: "I answered yes sometimes. If I had already found some nodes with cancer, it does not matter for the treatment of the patient if I find one more, then I answered no." (P6).

**4.4.3 Workflow and control.** All participants preferred using the tool over manual review, but the reasons varied. The junior pathologists experienced a reduced effort: "It wasn't as demanding. I could focus my energy and use it for smarter things. It was more fun" (P5). Senior pathologists preferred only the section navigation and counting aids "It will make my work faster. But I don't need the regions, it doesn't need to identify cancer. That it can give me the lymph nodes, so I don't have to navigate, that's enough for me." (P6).

They explicitly mentioned viewing the lymph node in **overview first** with **retained manual control** as essential. **Counting** was also helpful, with some even identifying manual counting as a risk to patient safety. They also expressed a preference for having the tool only when they opened it, giving them the option when to use it at all.

**4.4.4 Algorithms in the loop.** In order to facilitate discussions on automation issues, we showed participants a number of rejected prototypes. When shown an alternative prototype that highlighted the most positive regions by a colour-coded border, to our surprise, most expressed a preference for that alternative. When explicitly asked if this could lead to confirmation bias or risk of over-exploitation, most cited their work ethic and the vigilance required by their profession as natural barriers to such behaviours; *“If I find positives, the patient receives chemotherapy, and if I do not they are considered done. Finding these cancers are really important. I don’t think any pathologist would do that [over-reliance] just to get a longer lunch.”* (P4).

While many deemed that it would save time to not always show regions by removing some clear negatives, most also suggested that they or their environments would not be ready for such high-automation measures; *“It is tempting, but I don’t think we are ready for that. It needs to show one or two [suggested regions] just to make me feel confident.”* (P3).

We did not receive a clear indication of whether the interface could cause out of the loop issues, such as novice pathologists never learning the more efficient strategies of seniors. Some thought it could cause such issues, some thought it would not, and others said it would not matter since becoming dependant on their tools is part of the profession.

**4.4.5 Toward clinical use.** We asked participants whether they believed RAVS could become part of their everyday workflow. All participants suggested that personal experiences of RAVS in an experimental setting could not, on its own, convince them that the system was ready for clinical adoption. In addition to retrospective clinical studies and the required regulatory approvals, pathologists would want to perform local validation studies before any adoption. For example: *“You would need a large clinical study and local validations, but having done that, I would feel confident using such a system. Especially for these monotonous tasks like looking for lymph node metastases, that you know, many pathologists find boring.”* (P2). Additionally, any AI-system specialised for one specific task risk being unusable due to the fact that pathologists often need to identify *other* rare phenomena if they appear. The fact that sections were always viewed in overview first and that four regions were always presented made most participants confident that this would not be a barrier for RAVS when applied to colorectal lymph nodes. However, some participants still expressed some doubts: *“I am less worried about missing accidental findings after the experiment, but I still just think it is scary to start to lean on the algorithm so much.”* (P5).

## 5 EVALUATION: LEARNING FROM USE

We wanted to know if usage logging of individual region decisions could help further improve model performance when **learning**

**from use**, leading to an increased capability over time. In the short term, an increased capability might motivate using a lower sensitivity setting with RAVS and thus shorter review times. In the long term, it might mean redesigning the human-AI interaction altogether.

### 5.1 Data and Method

We used the 14 cases from the user study with the consensus labels as new training data, combined with the previous AIDA-LNCO [22] training data in a 10:90 ratio. As a test set, we randomly sampled seven cases from the AIDA-LNCO2 [23] dataset that had not been part of the user study. We then retrained the model using the same method and architecture as the previous model. One pathologist provided ground truth for the new test set. We then scored both the retrained model and the original model on the new test set.

### 5.2 Results

On the test set, the retrained model achieved a 99.9% AUC and perfect sensitivity at a 0.6% false-positive rate, which was an improvement over the original that achieved a 98.8% AUC and required a 10% false-positive rate for perfect sensitivity. Note that the quality is still at a sub-pathologist level. The relative improvement shows that improvement by learning as a side-effect of regular use was feasible in our experimental setting.

## 6 DISCUSSION

Through RAVS, we have shown that consideration of well-known HCI principles and AI-aware features allowed domain experts to work faster with maintained quality, despite the AI model performing below expert-levels. Our system demonstrates that conceptualising an automation-control balance along a single dimension is not always adequate. Rather, multiple features combine into an overall user experience with nuances and complex interconnections. We believe that an appropriate designerly mindset for creating successful human-AI interaction entails placing less emphasis on isolated visualisations. Instead, we call for more work that explore efficient collaboration strategies that consider and describe human, algorithmic, and most importantly, human-AI ensemble factors.

We designed RAVS so that the effective strategy to exploit AI predictions is so easy to use it cannot be missed. However, manual control is always available. To confidently use the assistance, the user needs to develop enough trust in the system. Recent developments around explainable AI have placed emphasis on the intrinsic workings of models, as well as employing human explanatory strategies such as counterfactuals and similar examples. Our work suggests that effects that arise from repeated exposure over time, coupled with a variable level of control, could also be an effective source to building trust, albeit *progressively*. For instance, the AI-assistant presented by Steiner et al. visualised potentially positive regions above a certain prediction threshold. They observed improved efficiency for target-present cases, but only minor improvements occurred on target-absent cases [33]. Based on the experiences in our work, we argue that not showing any regions for normal cases could be the cause of that non-improvement. Without being faced with normal regions, the users would not be able to explicitly learn about AI imperfections in target-absent cases, and thus distrusting

the result in those cases, reverting to manual methods. Our feature of **Always present regions** was designed to facilitate fast reviews of target-absent cases, but whether it is the main cause for the efficiency improvement achieved is difficult to say. Adding some validity to this speculation, Gu et al. [16] independently makes similar recommendations on always presenting regions for negative cases, based on formative evaluations of a comparable tool.

One shortcoming of our RAVS system was that it provided only limited support for users to distinguish malfunctions. Although this seemed to have little effect in our experimental condition, in clinical use we might observe less benefit of automatic support due to this ambiguity. Our feature of **Hide underlying probabilities** was identified as a possible barrier for becoming confident in identifying error characteristics. Providing the relative probabilities could be a solution that would still employ repeated exposure to build trust and model understanding. However, it is unclear whether such a change would create other issues, such as confirmation bias, overreliance on the strong positives or causing users to spend too much time worrying about the specific probability assigned. Interestingly, in the case where users noted that the model had placed suggested regions “where there can be nothing”, prediction heatmaps revealed problems, with the model assigning positive values on almost all grey pixels. These images lacked tumours, and whether a true tumour would have yielded higher relative values remains unknown.

Collecting training datasets and building trust with users are all daunting tasks to do upfront, especially if the targeted scenario leaves humans out of some decisions. An assistive model in the loop of human decision making might move this effort from an upfront task to being iteratively achieved in use. Our RAVS system demonstrates that a relatively low-effort ML model coupled with a user interface can be the source of workflow changes that in turn allow collecting large amounts of relevant training data that would be hard to achieve otherwise.

We applied RAVS to searching for cancer in colorectal lymph nodes. Besides being a prime candidate to work for lymph nodes from other organs (such as breast) and for other pathology applications, it could also have uses in other domains with high-resolution zoomable images such as scanning aerial photographs. The seven human-AI interaction features presented will probably not be applicable to all human-AI interaction designs. However, we do believe they can be generative for designers, allowing them to frame their own challenges in new ways.

Limitations of our study include the small number of pathologists and the imbalance of positive to negative sections stemming from our decision to focus on a clinically representative dataset. The focus on creating a system that would improve either efficiency or quality also prevented us from considering features in isolation and assessing their individual effects. There are also algorithmic challenges in learning from use. In our experiment, we used the consensus labels of the six pathologists to label training data. In a real setting, there would typically only be one pathologist diagnosing a case, leading to higher label noise.

## 7 CONCLUSION

We presented Rapid Assisted Visual Search, a system that assists pathologists in making fast high-quality decision by exploiting information from an imperfect AI model. We provided rationale for designers, identifying how HCI principles combine with seven AI features to give an overall user experience. Our evaluations show that our system is fast, accurate, helpful and can learn from use. The datasets used are available for extending our work.

## ACKNOWLEDGMENTS

We wish to thank Gordan Maras, M.D., without whom this project would not have been possible. The study was financially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP).

## 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] Péter Bárdi, Oscar Geessink, Quirine Manson, Marcory Van Dijk, Maschenka Balkenhol, Meyke Hermesen, Babak Ehteshami Bejnordi, Byungjae Lee, Kyunghyun Paeng, Aoxiao Zhong, Quanzheng Li, Farhad Ghazvinian Zanjani, Svitlana Zinger, Keisuke Fukuta, Daisuke Komura, Vlado Ovtcharov, Shenghua Cheng, Shaoqun Zeng, Jeppe Thagaard, Anders B. Dahl, Huangjing Lin, Hao Chen, Ludwig Jacobsson, Martin Hedlund, Melih Çetin, Eren Halıcı, Hunter Jackson, Richard Chen, Fabian Both, Jörg Franke, Heidi Küsters-Vandeveld, Willem Vreuls, Peter Bult, Bram van Ginneken, Jeroen van der Laak, and Geert Litjens. 2019. From Detection of Individual Metastases to Classification of Lymph Node Status at the Patient Level: The CAMELYON17 Challenge. *IEEE Transactions on Medical Imaging* 38, 2 (Feb. 2019), 550–560. <https://doi.org/10.1109/TMI.2018.2867350>
- [3] 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>
- [4] Wouter Bulten, Maschenka Balkenhol, Jean-Joël Awoumou Belinga, Américo Brilhante, Ashi 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, Hester van Boven, Robert Vink, Jeroen van der Laak, Christina Hulsbergen-van der Kaa, and Geert Litjens. 2020. Artificial Intelligence Assistance Significantly Improves Gleason Grading of Prostate Biopsies by Pathologists. *Modern Pathology* (Aug. 2020), 1–12. <https://doi.org/10.1038/s41379-020-0640-y>
- [5] Wouter Bulten, Hans Pinckaers, Hester van Boven, Robert Vink, Thomas de Bel, Bram van Ginneken, Jeroen van der Laak, Christina Hulsbergen-van de Kaa, and Geert Litjens. 2020. Automated Deep-Learning System for Gleason Grading of Prostate Cancer Using Biopsies: A Diagnostic Study. *The Lancet Oncology* 21, 2 (Feb. 2020), 233–241. [https://doi.org/10.1016/S1470-2045\(19\)30739-9](https://doi.org/10.1016/S1470-2045(19)30739-9)
- [6] 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>
- [7] Carrie J. Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. 2019. “Hello AI”: Uncovering the Onboarding Needs of Medical Practitioners for Human-AI Collaborative Decision-Making. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (Nov. 2019), 104:1–104:24. <https://doi.org/10.1145/3359206>
- [8] Gabriele Campanella, Matthew G. Hanna, Luke Geneslaw, Allen Mirafior, 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 (July 2019), 1301–1309. <https://doi.org/10.1038/s41591-019-0508-1>
- [9] Wild Christopher P. Weiderpass Elisabete, and Stewart Bernard W (Eds.). 2020. *World Cancer Report: Cancer Research for Cancer Prevention*. International Agency

- for Research on Cancer, Lyon, France.
- [10] C. C. Compton, L. P. Fielding, L. J. Burgart, B. Conley, H. S. Cooper, S. R. Hamilton, M. E. Hammond, D. E. Henson, R. V. Hutter, R. B. Nagle, M. L. Nielsen, D. J. Sargent, C. R. Taylor, M. Welton, and C. Willett. 2000. Prognostic Factors in Colorectal Cancer. College of American Pathologists Consensus Statement 1999. *Archives of Pathology & Laboratory Medicine* 124, 7 (July 2000), 979–994. [https://doi.org/10.1043/0003-9985\(2000\)124<0979:PFICC>2.0.CO;2](https://doi.org/10.1043/0003-9985(2000)124<0979:PFICC>2.0.CO;2)
- [11] Graham Dove, Kim Halskov, Jodi Forlizzi, and John Zimmerman. 2017. UX Design Innovation: Challenges for Working with Machine Learning as a Design Material. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, CO, USA) (CHI '17). Association for Computing Machinery, New York, NY, USA, 278–288. <https://doi.org/10.1145/3025453.3025739>
- [12] Mica R. Endsley. 2018. Level of Automation Forms a Key Aspect of Autonomy Design. *Journal of Cognitive Engineering and Decision Making* 12, 1 (March 2018), 29–34. <https://doi.org/10.1177/155534317723432>
- [13] Mathias S. Fleck and Stephen R. Mitroff. 2007. Rare Targets Are Rarely Missed in Correctable Search. *Psychological Science* 18, 11 (Nov. 2007), 943–947. <https://doi.org/10.1111/j.1467-9280.2007.02006.x>
- [14] Clifton Forlines and Ravin Balakrishnan. 2009. Improving Visual Search with Image Segmentation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Boston, MA, USA) (CHI '09). Association for Computing Machinery, New York, NY, USA, 1093–1102. <https://doi.org/10.1145/1518701.1518868>
- [15] Ben Green and Yiling Chen. 2019. The Principles and Limits of Algorithm-in-the-Loop Decision Making. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (Nov. 2019), 50:1–50:24. <https://doi.org/10.1145/3359152>
- [16] Hongyan Gu, Jingbin Huang, Lauren Hung, and Xiang 'Anthony' Chen. 2021. Lessons Learned from Designing an AI-Enabled Diagnosis Tool for Pathologists. arXiv:2006.12695 <http://arxiv.org/abs/2006.12695>
- [17] Todd Kulesza, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. 2015. Principles of Explanatory Debugging to Personalize Interactive Machine Learning. In *Proceedings of the 20th International Conference on Intelligent User Interfaces* (Atlanta, GA, USA) (IUI '15). Association for Computing Machinery, New York, NY, USA, 126–137. <https://doi.org/10.1145/2678025.2701399>
- [18] Ah-Young Kwon, Ha Young Park, Jiyeon Hyeon, Seok Jin Nam, Seok Won Kim, Jeong Eon Lee, Jong-Han Yu, Se Kyung Lee, Soo Youn Cho, and Eun Yoon Cho. 2019. Practical Approaches to Automated Digital Image Analysis of Ki-67 Labeling Index in 997 Breast Carcinomas and Causes of Discordance with Visual Assessment. *PLOS ONE* 14, 2 (Feb. 2019), 1–13. <https://doi.org/10.1371/journal.pone.0212309>
- [19] Martin Lindvall, Jesper Molin, and Jonas Löwgren. 2018. From Machine Learning to Machine Teaching: The Importance of UX. *Interactions* 25, 6 (2018), 52–57. <https://doi.org/10.1145/3282860>
- [20] Geert Litjens, Peter Bandi, Babak Ehteshami Bejnordi, Oscar Geessink, Maschenka Balkenhol, Peter Bult, Altuna Halilovic, Meyke Hermens, 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 (June 2018), giy065. <https://doi.org/10.1093/gigascience/gyi065>
- [21] Yun Liu, Krishna Gadepalli, Mohammad Norouzi, George E. Dahl, Timo Kohlberger, Aleksey Boyko, Subhashini Venugopalan, Aleksei Timofeev, Philip Q. Nelson, Greg S. Corrado, Jason D. HIPP, Lily Peng, and Martin C. Stumpe. 2017. Detecting Cancer Metastases on Gigapixel Pathology Images. arXiv:1703.02442 [cs]
- [22] Gordan Maras, Martin Lindvall, and Claes Lundström. 2019. *Regional Lymph Node Metastasis in Colon Adenocarcinoma*. <https://doi.org/10.23698/aida/Inco>
- [23] Gordan Maras, Martin Lindvall, and Claes Lundström. 2020. *Regional Lymph Node Metastasis in Colon Adenocarcinoma, Second Collection Series*. <https://doi.org/10.23698/aida/Inco2>
- [24] Luke Oakden-Rayner, Jared Dunnmon, Gustavo Carneiro, and Christopher Ré. 2019. Hidden Stratification Causes Clinically Meaningful Failures in Machine Learning for Medical Imaging. arXiv:1909.12475 [cs, stat]
- [25] Gabriel Pereyra, George Tucker, Jan Chorowski, Łukasz Kaiser, and Geoffrey Hinton. 2017. Regularizing Neural Networks by Penalizing Confident Output Distributions. arXiv:1701.06548 [cs]
- [26] Milda Pocevičiūtė, Gabriel Eilertsen, and Claes Lundström. 2020. Survey of XAI in Digital Pathology. In *Artificial Intelligence and Machine Learning for Digital Pathology: State-of-the-Art and Future Challenges*, Andreas Holzinger, Randy Goebel, Michael Mengel, and Heimo Müller (Eds.). Springer International Publishing, Cham, 56–88. [https://doi.org/10.1007/978-3-030-50402-1\\_4](https://doi.org/10.1007/978-3-030-50402-1_4)
- [27] Joaquin Quiñero-Candela (Ed.). 2009. *Dataset Shift in Machine Learning*. MIT Press, Cambridge, Mass.
- [28] Maithra Raghu, Katy Blumer, Greg Corrado, Jon Kleinberg, Ziad Obermeyer, and Sendhil Mullainathan. 2019. The Algorithmic Automation Problem: Prediction, Triage, and Human Effort. arXiv:1903.12220 [cs]
- [29] Ben Shneiderman. 1996. The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations. In *Proceedings 1996 IEEE Symposium on Visual Languages* (Boulder, CO, USA). IEEE, New York, NY, USA, 336–343. <https://doi.org/10.1109/VL.1996.545307>
- [30] Ben Shneiderman. 2020. Human-Centered Artificial Intelligence: Reliable, Safe & Trustworthy. *International Journal of Human-Computer Interaction* 36, 6 (April 2020), 495–504. <https://doi.org/10.1080/10447318.2020.1741118>
- [31] Robert Spence. 2002. Rapid, Serial and Visual: A Presentation Technique with Potential. *Information Visualization* 1, 1 (March 2002), 13–19. <https://doi.org/10.1057/palgrave/ivs/9500008>
- [32] Karin Stacke, Gabriel Eilertsen, Jonas Unger, and Claes Lundström. 2019. A Closer Look at Domain Shift for Deep Learning in Histopathology. arXiv:1909.11575 [cs]
- [33] David F. Steiner, Robert MacDonald, Yun Liu, Peter Truszkowski, Jason D. Hipp, Christopher Gammage, Florence Thng, Lily Peng, and Martin C. Stumpe. 2018. Impact of Deep Learning Assistance on the Histopathologic Review of Lymph Nodes for Metastatic Breast Cancer. *The American Journal of Surgical Pathology* 42, 12 (Dec. 2018), 1636. <https://doi.org/10.1097/PAS.0000000000001151>
- [34] Peter Ström, Kimmo Kartasalo, Henrik Olsson, Leslie Solorzano, Brett Delahunt, Daniel M Berney, David G Bostwick, Andrew J Evans, David J Grignon, Peter A Humphrey, Kenneth A Iczkowski, James G Kench, Glen Kristiansen, Theodor H van der Kwast, Katia R M Leite, Jesse K McKenney, Jon Oxley, Chin-Chen Pan, Hemamali Samarasinghe, John R Srigley, Hiroyuki Takahashi, Toyonori Tsuzuki, Murali Varma, Ming Zhou, Johan Lindberg, Cecilia Lindskog, Pekka Ruusuvaari, Carolina Wählby, Henrik Grönberg, Mattias Rantalainen, Lars Egevad, and Martin Eklund. 2020. Artificial Intelligence for Diagnosis and Grading of Prostate Cancer in Biopsies: A Population-Based, Diagnostic Study. *The Lancet Oncology* 21, 2 (Feb. 2020), 222–232. [https://doi.org/10.1016/S1470-2045\(19\)30738-7](https://doi.org/10.1016/S1470-2045(19)30738-7)
- [35] Sten Thorstenson, Jesper Molin, and Claes Lundström. 2014. Implementation of Large-Scale Routine Diagnostics Using Whole Slide Imaging in Sweden: Digital Pathology Experiences 2006–2013. *Journal of Pathology Informatics* 5, 1 (March 2014), 14. <https://doi.org/10.4103/2153-3539.129452>
- [36] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y. Lim. 2019. Designing Theory-Driven User-Centric Explainable AI. 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, 1–15. <https://doi.org/10.1145/3290605.3300831>
- [37] Qian Yang. 2018. *Machine Learning as a UX Design Material: How Can We Imagine Beyond Automation, Recommenders, and Reminders?* AAAI Spring Symposium Series Technical Report SS-18. Stanford. 6 pages. <https://aaai.org/ocs/index.php/SSS/SSS18/paper/view/17471>
- [38] Qian Yang, Alex Scuito, John Zimmerman, Jodi Forlizzi, and Aaron Steinfeld. 2018. Investigating How Experienced UX Designers Effectively Work with Machine Learning. In *Proceedings of the 2018 Designing Interactive Systems Conference* (Hong Kong, China, 2018) (DIS '18). Association for Computing Machinery, New York, NY, USA, 585–596. <https://doi.org/10.1145/3196709.3196730>
- [39] Qian Yang, Aaron Steinfeld, Carolyn Rosé, and John Zimmerman. 2020. Re-Examining Whether, Why, and How Human-AI Interaction Is Uniquely Difficult to Design. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '20). Association for Computing Machinery, New York, NY, USA, 1–13. <https://doi.org/10.1145/3313831.3376301>
- [40] Qian Yang, John Zimmerman, Aaron Steinfeld, and Anthony Tamasic. 2016. Planning Adaptive Mobile Experience s When Wireframing. In *Proceedings of the 2016 ACM Conference on Designing Interactive Systems* (Brisbane, QLD, Australia, 2016) (DIS '16). Association for Computing Machinery, New York, NY, USA, 565–576. <https://doi.org/10.1145/2901790.2901858>