Domain Expert in the Loop of Digitized Histopathology Education and Artificial Intelligence

Erika Váczlavová*
Supervised by: Miroslav Laco†

Faculty of Informatics and Information Technologies
Slovak University of Technology
Bratislava, Slovakia

Abstract

In this paper, we propose a way to use a graphical user interface to present digitized multi-modal data in the field of medicine for specific domain experts. Our data consisted of digitized histopathology specimens, subject to expert examination. As the digitization of histopathology for educational purposes is only in its beginning stages, we explore how to present the data to experts in a way to encourages them to build up their confidence in digitized workflow. As part of this research, we are working on streamlining the workflow by designing assistance tools based on artificial intelligence (AI). While presenting the results of AI to specific domain experts in medicine, it is important to choose the right explainability of the results of black-box algorithms, and how to present the outputs in the user interface. We found out that the implementation of functionalities driven by artificial intelligence depends on the level of expertise of the domain expert. The differences are observed in a case study with cooperation from medical students and doctors, who got access to digitized multi-modal data with AI-powered functionalities in iteratively designed prototypes of the specialized system for education in the field of histopathology. We present outcomes from the aforementioned case study to serve as a base for the future development of specialized interfaces in the field of digitized histopathology.

Keywords: Histopathology, User Experience, Artificial Intelligence

1 Introduction

Histopathological specimens are samples obtained through biopsy or surgical procedures and subjected to histological processing. Histological processing involves the fixation of the sample, cutting it into thin sections, and staining with specific dyes that allow for microscopic tissue analysis. To digitize these glass specimens are used special scanners with whole slide imaging technology (WSI). WSI produces high-resolution digital images at multiple magnifications and focal planes. These types of images are highly suitable for educational purposes as the WSI is more interactive, it is easy to share them, and provides the opportunity to convey the same information to each student, which is not possible with glass slides, because none of them are identical. Hence, it is not surprising that WSI is increasingly being used in examinations[8].

During the examination of a slide, pathologists carefully observe and interpret the histological characteristics of the case within the context of clinical information. Through this process, they identify regions of interest, that are pertinent to the specific cases[10]. The whole process of an examination of slides and annotation is time-consuming and inefficient because areas of interest cannot be marked directly into glass slides and to determine the area of interest the specialist must go through the whole specimen in multiple zoom views.

Higher accuracy, capability, and efficiency are some of the many reasons why to transform the workflow to a digital one, through the digitization of specimens. By digitization, WSI images replace the glass slides. These WSI images are accessible by annotation tools provided in a digital platform. These tools typically provide a menu of markup shapes including measured lines, polygons, rectangles, circles, and free-form lines, which can be applied in a wide range of colors. Some systems allow text labeling of the annotation[10].

Another method to enhance the efficiency of histopathologist's work in annotating individual WSI images is by integrating AI algorithms into the process. These algorithms can automatically identify areas of interest within the images using different approaches, thereby accelerating the workflow of experts. Subsequently, experts would review the outputs of the artificial intelligence system and make adjustments as needed.

We aim to leverage the benefits of digitized image annotation processes into the teaching process at medical universities. Our endeavor involves developing a specialized tool equipped with diverse educational features, and functionalities supported by AI to the extent that its

*xvaczlavova@stuba.sk
†miroslav.laco@stuba.sk

results are presented according to the target audience.

2 Human-in-the-loop of Artificial Intelligence

Human-in-the-loop Artificial Intelligence (HITL), refers to a process addressing concerns of individuals regarding the negative impacts of the artificial intelligence revolution, such as output accuracy and interpretability. This process integrates the operation of artificial intelligence with the human factor based on domain knowledge. In the case of supervised learning, the AI can learn and make decisions based only on the supplied data, along with tags associated with the data, which we call annotations. AI's decisions are based on statistics and connections, abstracted at both lower and higher levels from the supplied data. But these decisions do not contain domain knowledge, which often does not appear in the data [4]. While training models of AI, human input is often important, which corrects the results and thereby helps improve algorithms. HITL also addresses ethical questions about ownership of knowledge on which artificial intelligence models are trained since the models they learn from data created by ordinary workers [13].

3 Artificial Intelligence and User Experience

Artificial Intelligence (AI) holds a pivotal role in improving human-computer interaction and optimizing user experience. However, the design and innovation of such interactions pose multifaceted challenges. AI's potential for introducing unforeseen errors can adversely impact both reputation and user experience in collaborative settings. Designing cooperation between humans and AI is particularly demanding [12].

In iterative prototyping and testing of user experience without the use of AI, it is possible to address and test further iterations of shortcomings. However, when prototyping and testing with AI features, this becomes challenging as the AI may introduce unforeseen errors. Another challenge for designers is setting user expectations regarding what can be expected from the AI. Since the AI lacks legal and ethical awareness, there is concern over incorrect outputs potentially causing frustration. Additionally, for user experience professionals, collaborating with artificial intelligence experts can be challenging due to the distinct domains involved. Moreover, by prioritizing explainability in AI, users can develop a deeper trust in the system, as it allows them to comprehend the inner workings and decision-making processes, thereby ensuring an optimal balance of complexity in presented results. [12, 5].

Various methods exist for presenting AI model output data. Designers must consider scenarios like true positives, false positives, true negatives, and false negatives. These are addressed in two result generation approaches. One prioritizes output precision, aiming for accuracy even if the output set is smaller, potentially overlooking some true positives. The other approach, called recall, aims for a broader set of outputs to maximize the presence of true positives, even if not all results are relevant or correct.[2].

4 State of the Art in Education Process

We focused on analyzing various educational and annotation tools for digitized multi-modal data that can be used in both teaching and practice in medical universities. In these tools, we look at functionalities that are useful in the study of pathology, as well as in the analysis of medical image data. In the realm of medical imaging and education, several tools have emerged, each with its own set of advantages and limitations.

QuPath stands as a platform for the analysis of medical image data. Its ability to handle diverse formats and provide a range of marking tools empowers users to annotate and manipulate areas of interest directly onto digital specimens. However, the absence of a comment feature and a somewhat complex user interface may pose challenges, particularly for those with limited computer literacy. In contrast, **AMBOSS** represents a commercially driven approach, offering a repository of educational materials in a sleek, user-friendly interface. Its virtual library and note-taking functionalities enhance the learning experience, allowing users to create and share annotations with ease. Nonetheless, its closed nature restricts the ability to modify or expand the content of medical knowledge for people in medical field study. Meanwhile, **The Human Protein Atlas** serves as a valuable supplementary resource, providing a wealth of high-resolution images showcasing protein distribution across various human tissues and cell lines. While its predefined pathways and detailed descriptions offer structured learning experiences, the inability to insert custom images or annotate specimens may restrict its utility for interactive study. In essence, each tool brings unique strengths to the table. However, navigating their respective limitations is crucial in harnessing their full potential for medical education and research in the modern era.

5 Our Approach

Recent studies recommend that for working with data in the field of medicine and health, it is necessary to develop new usability methods and theories on how to work with them [6]. Based on these recommendations, various new procedures began to emerge as to how to

perceive the user during the design of the system and also that this user needs to be specified more closely according to the domain area in which the design is being created. One viewpoint entails the adaptation of the conventional user-centered design principle, particularly within the medical domain, where it has been redefined as patient-centered design[6]. Based on the state-of-the-art in patient-centered design, we decided to modify this principle and work on histopathology-expert-centered design and medic-centered design.

5.1 Design Centered on the Domain Expert in the Field of Medicine

Before creating a functional system design, it is essential to consider all stakeholders involved in the creation process, ensuring that the trust of the domain expert, for whom the system is designed, is gradually established. All of these stakeholders are visualized in Fig.1. Additionally, the system should be designed in such a way that the domain expert can naturally utilize all its functionalities and extensions without hesitation after its creation.

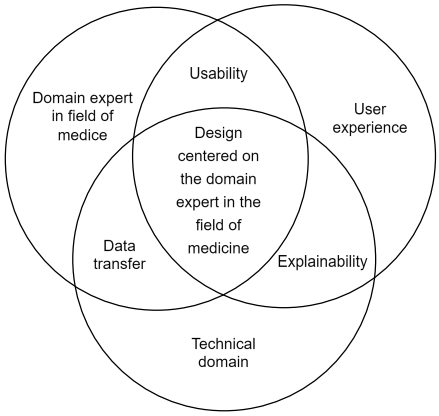

Figure 1: Visualization of 3 components and their cooperation in design centered on the domain expert in the field of medicine methodology. Each area overlay represents an area to focus on. Adapted from Meloncon et al. [6]

During the system development process, the role of this principle is to ensure that the domain expert generates data required for the technical aspects of the system, while the system provides data to the domain expert in an understandable format. The presentation format of the data is examined by a user experience expert, who explores how to create a reliable and usable system. Findings are obtained through interaction with the domain expert and translated into technical language for the development.
To create a design focused on the domain expert in medicine, it is important to build the entire collaboration thoroughly and approach the design as a continuously evolving relationship between stakeholders. To be able to design the cooperation and the system to the satisfaction of the domain expert, it is important to get the collaboration right from the initial stages steered properly.

5.2 Annotation Enhanced Educational Tool

As a second important contribution of this paper, we have designed an annotation tool that will be part of a comprehensive educational system in cooperation with domain experts from the medical university. The proposed prototype focused on functionalities related to annotating digitized histopathological specimens. In this prototype, individual images can be viewed and annotated using various tools. These tools are divided into those not supported by AI and those simulating real results of AI models.
The design was based on user needs of real users, which we generalized into 2 personas. These personas served to better understand the mental model of end users. Among these personas is an expert who teaches histology and pathology at the university and practices in the clinical sphere, aims to teach modern methods at the university, and provides feedback to students on their work while also sharing extra materials. The second persona is a student who seeks hands-on experience in annotation and desires access to materials even after classes to further educate themselves in the field of diagnosis determination.

5.3 Presentation of the Artificial Intelligence Outputs in Histopathology

While designing the user interface for the annotation tool, our focus was on deliberating upon the most suitable presentation of artificial intelligence. Two principal approaches were considered: automation characterized by AI-driven task execution devoid of human intervention, and augmentation which entailed AI providing recommendations to users of the annotation tool, who subsequently validated or dismissed its outputs within the context of our work[9].

We proposed three functionalities aimed at simulating AI results in various forms. Automation was represented by a tool that upon triggering the workflow, automatically highlighted all areas of interest on the annotated image. Augmentation was depicted through two tools: one gradually revealed areas of interest in the annotated image, requiring user confirmation or rejection with each annotation. The second tool offered the option of displaying hints, outlining regions on the image where areas of interest could potentially be found, without showing the actual annotation. Our contribution includes comparing the usability and the explainability of AI outputs using user experience methods such as usability testing and contextual interviews.

5.4 Reward Function and Explainability

A reward function is a component of reinforcement learning of AI algorithms. It defines the objective or goal that the AI is trying to maximize or minimize in order to receive rewards or punishments, guiding the AI's behavior toward achieving desired outcomes. This approach may be considered from the UX point of view when working with all NNtypes, not only reinforcement-learning-based NNs.

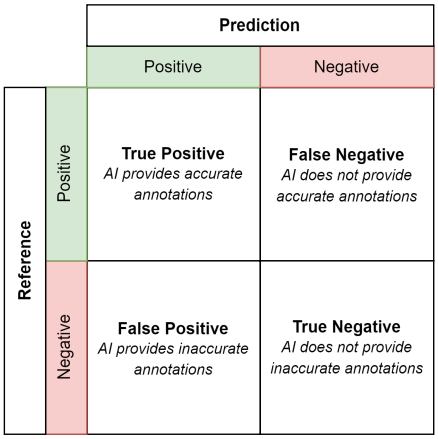

Figure 2: Reward function for AI used in annotation tool.

In Figure 2, errors arise in two cases. False Positives occur when AI provides inaccurate annotations, especially troubling in fields like education or medicine due to potential user impact. False Negatives happen when AI fails to provide accurate annotations, leading to increased manual work for users and posing challenges in maintaining focus during correction. We consider this as a concern for applications in the field of medicine where it is crucial for the user to receive accurate outputs. Therefore, our proposal offers the user more benefits in minimizing False Positives (where AI generates inaccurate annotations), hence it is appropriate to optimize the function for precision. We acknowledge that opting for this method entails a compromise, meaning our model will lean more towards abstaining from creating any annotation rather than producing an inaccurate one. We prioritize refraining from displaying any annotation to the user over presenting an inaccurate one. We propose to verify this approach and evaluate our proposal using pre-generated annotations as AI's results against manually created annotations.

6 Approach Validation and Testing

Based on contextual inquiry and observation of domain experts we prepared four user scenarios, which were, according to the good practice in the field of UX and usability testing, tested by five participants (more in [7]). These participants corresponded to the personas created in earlier stages of the project. The group of participants consisted of four students with low expertise and one experienced expert from the medical field. Despite varying IT skills, all possess sufficient knowledge to effectively annotate the data.
The usability testing was focused on evaluating the necessity and form of implementing AI. We measured quantitative outcomes such as task completion rate, error rate, and time needed for tasks. We also collected qualitative data from user feedback and suggestions. During this testing phase, we also examined the explainability of artificial intelligence outcomes and how to present them to end users, as represented by the testers. The entire testing process was conducted using the thinking-aloud method [1].

We defined four tasks. The tasks were created in cooperation with the domain expert. One of them was designed for the user not to utilize tools supported by AI. The remaining tasks simulated various ways of utilizing the outcomes of AI. The outcomes from AI were simulated by pre-created annotations, created by domain experts, and served to participants using the Wizard of Oz methodology [3, 11]. All tasks were about annotation in real digitized specimen. The specimen was from cardiac tissue and contains the endocardium. All tasks were based on the same annotations on the same data. The tasks were:

1. Please annotate the endocardium in the image using the drawing function - This task was designed because its results will serve as a baseline for evaluation.

2. Please annotate the endocardium in the image using the Annotation proposals function - This task was designed for observing the user behavior and the impact of augmentation on performance in simple tasks.

3. Please annotate the endocardium in the image using the Automated annotation function - This task was designed for observing the user behavior and the impact of automation on performance in simple tasks.

4. Please annotate the endocardium in the image using the Hints function - This task was designed for observing the user behavior and the impact of augmentation on performance in simple tasks.

7 Results

During the testing, we monitored various metrics, and after evaluation, we divided the results into quantitative and qualitative outcomes. Each relevant feedback obtained during testing helped us understand how to implement AI-supported features properly.

7.1 Quantitative Results

The typical procedure involves observing task completion success. In this case, participants were able to complete all tasks, with the assistance of a facilitator required only once. Task assignments were straightforward, and the prototype was designed to be trivial to users, as the end user is not proficient in information technology.

Task completion time

Participant (Expertise)	Manual annotation (Task 1)	Annotation proposal (Task 2)	Automated annotations (Task 3)	Hints (Task 4)
P1 (low)	145	98	70	**52**
P2 (low)	150	80	60	**40**
P3 (low)	180	96	**60**	69
P4(high)	185	75	50	**39**
P5 (low)	180	110	57	**42**
Average	168	91.8	59.4	**48.4**
Std	16.91	12.71	6.43	11.28

Table 1: Task completion time in seconds "Std" - Standard deviation

Time needed for one annotation

Participant (Expertise)	Manual annotation (Task1)	Annotation proposal (Task 2)	Automated annotations (Task 3)
P1 (low)	7	13	**6**
P2 (low)	8	7	**5**
P3 (low)	10	7	**5**
P4 (high)	12	**7**	10
P5(low)	11	10	**4**
Average	9.6	8.8	**6**
Std	2.07	2.68	2.10

Table 2: Time needed for one annotation in seconds;

In Table 1, we can observe the trend in user performance evolution within the proposed tool with the assistance of AI-generated results. As evidenced, tasks utilizing artificial intelligence were completed faster. These values must also consider a slight bias introduced by participants gaining experience with the tool and gradually acclimating to its use with each task. However, this bias is not high enough to preclude the assertion that the application of artificial intelligence in any form has increased work efficiency.

In Table 2, we can compare the time required for creating a single annotation in a digitized image. These results may be influenced by biases stemming from prototype limitations. However, this bias is not significant and applies to each annotation, so it does not need to be taken into consideration. In Table 2, it is evident that creating annotations without tools incorporating AI assistance takes longer than creating annotations with their use. Based on the numerical values, it is therefore most suitable to implement AI in the form of automation to reduce the time required for each annotation.

7.2 Qualitative Results

We conducted a qualitative evaluation based on participant observations during testing, and analysis of video recordings obtained during testing with the participant's consent to anonymously participate in the research, as well as the facilitator's questions or questionnaire inquiries.

Technical design of the prototype

All inquiries regarding the simplicity of application usage were responded to by participants with a positive sentiment. Considering that the prototype was designed so that participants meeting the parameters of our personas had no issues with its utilization, we deem it a suitable environment for testing the prototype with functions working with AI's results.

Automation

In the prototype, we represented automation through the functionality of displaying automatic annotations being simulated outputs of the AI for the given whole slide image instantly with the image itself as an overlay. All participants appreciated having a large number of annotations quickly using this approach. Regarding the facilitator's question about whether this functionality could pose any negative impact on their work or study, responses varied depending on the level of expertise.

Participants with lower levels of expertise stated that they appreciated such functionality as it speeds up their review of individual images during study or in their potential future work.

Participants with higher levels of expertise exhibit more skepticism towards this functionality. When using it, they are concerned that the system may offer incorrect annotations which they may not have time to verify and could potentially lead to errors. They also emphasize the importance, particularly in teaching contexts, of reviewing images to determine whether the area annotated is correct, which may not occur when a large number of annotations are displayed.

Augmentation

In the prototype, we represented augmentation of the manual annotation process with the AI outputs relevant to the given whole slide image through two different functionalities. One of them was Hints which displayed regions in the image where an area of interest might be located, prompting the creation of an annotation. Regardless of expertise, all participants appreciated this

functionality and claimed they would commonly use it.
Participants with lower levels of expertise would utilize this functionality during study sessions, where it would assist them in orienting themselves in the image and guiding them to create their own annotations.
Domain experts with higher levels of expertise liked this tool and claimed it would facilitate their manual and laborious examination of specimens. They also appreciated its potential for use in the educational process.

The second functionality representing augmentation is the Annotation proposal, where annotations gradually appear on the image, allowing the user to confirm or remove them with a click. After approval, they can continue to edit them. This functionality was perceived by all participants as faster than manual annotation but slower than automation. Regardless of expertise level, this functionality was perceived most positively, as domain experts felt they had control over individual annotations.

Augmentation or Automation
When asked which annotation functionality they would prefer, domain experts with lower levels of expertise agreed that automation seemed practical and fast. Conversely, domain experts with higher levels of expertise recommend augmentation, both in clinical practice and in teaching and study.

The level of trust in AI results
When asked whether they trust the system that recommends annotations, participants expressed skepticism. None of them confirmed that they would fully trust the system. However, this fact is positive because they would all verify the majority of annotations, thus reducing the risk of error.
The level of expertise of the domain expert, in this case, the participant, also influences their trust in the system. Participants with lower levels of expertise stated that if they had more knowledge in the respective field, they might be able to trust the system more. They also expressed that they would trust the system if they knew it didn't make errors frequently.
Participants with higher levels of expertise state that it's not possible to trust the tool 100 percent, but that's also true for humans. However, an individual who is still learning about the subject must have input on which to build. Such utilization of artificial intelligence would be credible only if an expert intervenes in the learning process to correct any misinformation provided to students. The use of such a system in clinical practice and trust in it would likely require time to build. The longer the tool is used, the more an expert would know which potential errors to focus on.
One of them stated:
"Even though I have the opportunity to intervene, the human mind tends to seek simpler paths. So, in that case, I wouldn't have trust in the system because ultimately I no longer trust myself. Comparing what I know with the information provided by the system can lead to a situation where two pieces of information confront each other. And now it's about which of those personalities is more confident to say 'but this is how it is,' even though it hasn't looked at, for example, 80,000 slides like artificial intelligence has. Because it will have a greater opportunity to feed its head than the human."

Explainability
During testing, we also asked our participants how their trust in the system could be supported. Their trust could be enhanced through explainability features, which would justify the individual outcomes of artificial intelligence. Explainability could assist domain experts in understanding highlighted areas and provide additional information necessary to confirm or decline AI results.
From the interviews, we learned that domain experts would prefer explainability in written form. This explainability should clarify the reasons why the annotation was created using medical terminology. Test participants did not prefer explainability in the form of percentages indicating AI's confidence in the annotation, nor did they favor heat-maps or other numerical ratings. Similarly, they did not prefer explainability in the form of a similar case shown in a tooltip.

8 Discussion

In the realm of the user experience in digitized histopathology, it is imperative to meticulously consider and accommodate the varying levels of expertise among domain-specific experts. This entails a conscientious approach to integrating the insights and contributions of experts from diverse domains, ensuring that each individual's specialized knowledge and proficiency are fully present the AI results in a proper way and leveraged to adapt the processes and outcomes within the digital histopathology framework.
Prioritizing expertise levels among domain experts in designing the frameworks for digital histopathology is fundamental for driving innovation and enhancing medicine study field improvements.

Assuming that the proposed tool will be utilized by experts with a lower level of expertise and also by experts with a high level of expertise, it is essential to design it in a manner that caters to the specific needs of both groups.
The limitation of the usability study was introduced by participants gaining experience with the tool and gradually acclimating to its use with each task. However, we claim this bias does not contradict the basal finding that the application of an AI-assisted approach in any form increases work efficiency when introduced in the tool after the user gets familiar with the manual annotation workflow.

8.1 Phases of Design Focused on the Domain Expert in the Field of Medicine

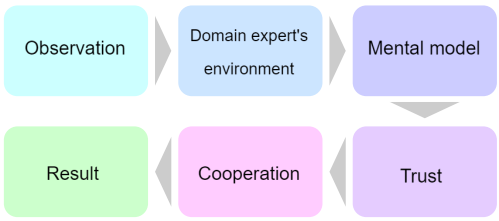

Figure 3: Visualization of phases of design focused on the domain Expert in the field of medicine.

Before creating a design focused on a domain expert in the field of medicine, it is important to build the entire collaboration thoroughly and approach the design as a continuously evolving relationship between stakeholders. To be able to design in a way that achieves satisfaction for the domain expert, it is important to grasp the collaboration correctly from the initial phases. All detailed steps of collaboration is depicted in Fig.3.

The first half of the collaboration ends with the creation and understanding of the mental model of the domain expert in the field of medicine. Only at this stage can we create specific personas and design to reference real user needs. This phase is preceded by observing domain experts. Observing domain experts also includes contextual interviews, obtaining a concrete picture of their needs. When it comes to a specific domain, such as in this case the domain of medicine, it is important to conduct observations and meetings with domain experts in their own working environment. Observing the home environment helps us understand the typical flow of activities, and the domain expert appears more confident in their familiar environment, with their behavioral model not being distorted by various external factors.

The second half of the collaboration begins with the most important phase, which is building trust. Trust from the domain expert in the field of medicine towards the technical domain expert is crucial due to significant differences in focuses. Trust needs to be built gradually through dialogue and openness. With such an approach, the domain expert gains confidence and begins to collaborate with technical domain experts as colleagues without the need to distinguish or underestimate either side. After gaining trust, it is necessary to reinforce it by involving the domain expert in the development process, making it clear that their opinion is important even in the technical domain. The involvement of the domain expert can take various forms, from the initial stage of prototype development during sketching, through testing or data creation, to feedback.

The final phase of the collaboration is the outcome. The outcome consists of multiple goals from each party involved in the collaboration. The outcome includes the developed prototype, product, and system, as well as the satisfaction of the domain expert in the field of medicine. The final phase may define various outcomes in different cases, but it is important for these outcomes to meet the goals and bring benefits to both domain experts in the field of medicine and the technical domain. Among the outcomes, we also include associated partial goals such as gained trust or expanded knowledge in the domain.

8.2 Application of Automation and Augmentation Based on the Level of Expertise in the Field of Medicine

It is crucial to focus on the implementation of intelligent features into tools used for educating domain experts in the field of medicine to streamline the work of medical professionals and generate a wealth of valuable study materials that will be more readily available to all students compared to current educational methods. Through collaboration, testing, and observation, we have found that how artificial intelligence is implemented into applications in the field of medicine should be on the expertise level of individual domain experts who will be using the proposed tools.

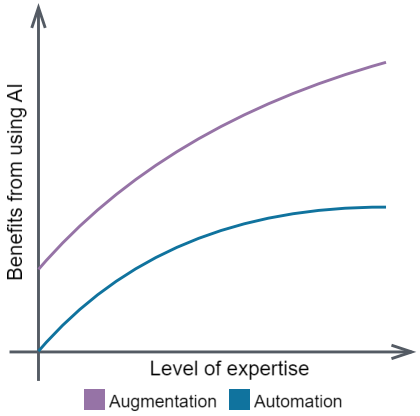

Figure 4: Visualization of the relationship between the benefits of using AI and levels of expertise in 2 types of AI implementation

There is a relationship between the expertise level of the domain expert and the number of benefits that can be derived from using AI-supported functionalities. As we can observe, the higher the expertise level, the greater the benefits provided by such functionalities. The reason is that experts with higher levels of expertise can critically evaluate the results of AI, whereas, without domain knowledge, there could be negative influences on the outcomes of AI.

As visualized in Fig.4, there are some differences between the methods of implementing AI and the benefits these methods yield to domain experts with varying levels

of expertise.
The first method of presenting AI outputs in the user interface, depicted in blue in the figure, is automation. This curve commences at the origin of the coordinate system, signifying that in the absence of domain knowledge, automation yields no discernible benefits either in the learning process or in clinical practice. This is because the user attempting to educate themselves through the application loses a crucial part of the learning process, namely analysis. The magnitude of benefits conferred by automation increases gradually with the accumulation of domain knowledge.

The second curve depicted in Fig.4, represented in purple, has its benefit value with near-zero domain knowledge obviously higher than the automation approach. As evident, this curve does start at a higher point, indicating that this method of AI implementation is suitable even in the educational process, where it can provide recommendations and help experts with lower levels of expertise navigate through digitized preparations. With increasing levels of expertise, the number of benefits that augmentation can bring also increases. In this method of AI implementation, domain experts have significant control over the AI results, which gives them a greater sense of comfort and increases trust.

9 Conclusion

In this paper, we targeted the identified research problem of domain experts in the loop of digitized histopathology education and artificial intelligence. We addressed these open research questions with our approach proposal including histopathology-expert-centered design and medic-centered design. We validated and examined the proposed approach in a case study in cooperation with domain experts from a medical university. Our main contribution is the phases of design focused on the domain expert in the field of medicine and the proposal of application of automation and augmentation based on the level of expertise in the field of medicine.
In our future work, we plan for medical faculty students to adopt the annotation tool in their education process as an extensive usability study while evaluating their interactions. The study on interactions will supplement the study on Design Focused on the Domain Expert in the field of medicine.

References

[1] Ted Boren and Judith Ramey. Thinking aloud: Reconciling theory and practice. *IEEE transactions on professional communication*, 43(3):261–278, 2000.

[2] Michael Buckland and Fredric Gey. The relationship between recall and precision. *Journal of the American society for information science*, 45(1):12–19, 1994.

[3] Steven Dow, Blair MacIntyre, Jaemin Lee, Christopher Oezbek, Jay David Bolter, and Maribeth Gandy. Wizard of oz support throughout an iterative design process. *IEEE Pervasive Computing*, 4(4):18–26, 2005.

[4] Fabrice Jotterand and Clara Bosco. Keeping the "human in the loop" in the age of artificial intelligence: accompanying commentary for "correcting the brain?" by rainey and erden. *Science and Engineering Ethics*, 26:2455–2460, 2020.

[5] Pantelis Linardatos, Vasilis Papastefanopoulos, and Sotiris Kotsiantis. Explainable ai: A review of machine learning interpretability methods. *Entropy*, 23(1):18, 2020.

[6] Lisa K Meloncon. Patient experience design: Expanding usability methodologies for healthcare. *Communication Design Quarterly Review*, 5(2):19–28, 2017.

[7] Jakob Nielsen and Thomas K Landauer. A mathematical model of the finding of usability problems. In *Proceedings of the INTERACT'93 and CHI'93 conference on Human factors in computing systems*, pages 206–213, 1993.

[8] Liron Pantanowitz. Digital images and the future of digital pathology. *Journal of pathology informatics*, 1, 2010.

[9] Sebastian Raisch and Sebastian Krakowski. Artificial intelligence and management: The automation–augmentation paradox. *Academy of management review*, 46(1):192–210, 2021.

[10] Chetan L Srinidhi, Seung Wook Kim, Fu-Der Chen, and Anne L Martel. Self-supervised driven consistency training for annotation efficient histopathology image analysis. *Medical Image Analysis*, 75:102256, 2022.

[11] Leandro Manuel Reis Velloso and Gil Barros. Recurrent techniques used in ux design: a report from user survey and interviews with professional designers. *Journal of Design Research*, 21(1):47–61, 2023.

[12] Qian Yang, Aaron Steinfeld, Carolyn Rosé, and John Zimmerman. Re-examining whether, why, and how human-ai interaction is uniquely difficult to design. In *Proceedings of the 2020 chi conference on human factors in computing systems*, pages 1–13, 2020.

[13] Fabio Massimo Zanzotto. Human-in-the-loop artificial intelligence. *Journal of Artificial Intelligence Research*, 64:243–252, 2019.