Linköping Studies in Science and Technology
Dissertations, No. 2157

Designing with Machine Learning in Digital Pathology

Augmenting Medical Specialists through Interaction Design

Martin Lindvall

Linköping Studies in Science and Technology
Dissertations, No. 2157

Designing with Machine Learning in Digital Pathology

Augmenting Medical Specialists through
Interaction Design

Martin Lindvall

Department of Science and Technology
Division of Media and Information Technology
SE-601 74 Norrköping, Sweden

Linköping, 2021

Designing with Machine Learning in Digital Pathology

Augmenting Medical Specialists through Interaction Design

MARTIN LINDVALL

ISBN 978-91-7929-604-9

ISSN 0345-7524

© MARTIN LINDVALL, 2021 (unless stated otherwise).

Published articles have been reprinted with permission from the respective copyright holder.

This work is licensed under CC BY-NC 4.0. To view a copy of this license, visit <http://creativecommons.org/licenses/by-nc/4.0/>

Cover: The figure depicts a lymph node from the AI-LNCO dataset that was produced as part of this work. The image is split into three parts, highlighting the role of data collection, ML model development, and interaction design, corresponding to the tripartite framework suggested in Chapter 7. The top part shows training data as blue lines, the middle part depicts the raw output of an ML model visualised as a heat map, and the bottom part shows regions suggested for manual review in a manner similar to the prototype on AI-assisted visual search described in Paper 2.

Typeset using L^AT_EX.

Printed by LiU-Tryck

Linköping, 2021

Abstract

Recent advancements in machine learning (ML) have led to a dramatic increase in AI capabilities for medical diagnostic tasks. Despite technical advances, developers of predictive AI models struggle to integrate their work into routine clinical workflows. Inefficient human-AI interactions, poor sociotechnical fit and a lack of interactive strategies for dealing with the imperfect nature of predictions are known factors contributing to this lack of adoption.

User-centred design methods are typically aimed at discovering and realising desirable qualities in use, pragmatically oriented around finding solutions despite the limitations of material- and human resources. However, existing methods often rely on designers possessing knowledge of suitable interactive metaphors and idioms, as well as skills in evaluating ideas through low-fidelity prototyping and rapid iteration methods—all of which are challenged by the data-driven nature of machine learning and the unpredictable outputs from AI models.

Using a constructive design research approach, my work explores how we might design systems with AI components that aid clinical decision-making in a human-centred and iterative fashion. Findings are derived from experiments and experiences from four exploratory projects conducted in collaboration with professional physicians, all aiming to probe this design space by producing novel interactive systems for or with ML components.

Contributions include identifying practical and theoretical design challenges, suggesting novel interaction strategies for human-AI collaboration, framing ML competence for designers and presenting empirical descriptions of conducted design processes. Specifically, this compilation thesis contains three works that address effective human-machine teaching and two works that address the challenge of designing interactions that afford successful decision-making despite the uncertainty and imperfections inherent in machine predictions. Finally, two works directly address design-researchers working with ML, arguing for a systematic approach to increase the repertoire available for theoretical annotation and understanding of the properties of ML as a designerly material.

Keywords: Human-AI interaction, interaction design, machine learning.

Populärvetenskaplig sammanfattning

Artificiell Intelligens (AI) har nyligen gjort stora framsteg inom medicinsk diagnostik. Trots de tekniska avancemangen så kvarstår stora utmaningar med att integrera resultaten i användbara verktyg för vardaglig klinisk verksamhet. Kända faktorer som försvårar en lyckad integration är ineffektiva människa-AI interaktioner, bristande anpassning till kliniska arbetsflöden och en avsaknad av strategier för att hantera imperfekta prediktioner.

Användarcentrerade designmetoder fokuserar typiskt på att upptäcka och realisera åtråvärda användningskvaliteer genom ett systematiskt iterativt arbete genom koncept, skisser och interaktiva prototyper. Existerande metoder är beroende av designerns föreställningsförmåga och kunskap kring befintliga interaktionsidiom och gränssnittsmetaforer, som tillåter en progressiv utdetaljerings av det tilltänkta användargränssnittet. Data-drivna maskininlärningsbaserade gränssnitt försvårar föreställningsförmågan. Tidiga konceptskisser och enkla prototyper som inte tar hänsyn till hur prediktioner beter sig riskerar att leda till stora svårigheter att anpassa algoritmer till ett koncept som inte är förankrat i hur tekniken fungerar eller inte beakta de specifika problem som imperfekta prediktioner ger upphov till.

Genom ett tillvägagångssätt förankrat i designforskning utforskar mitt arbete hur vi kan designa system med AI-komponenter som stödjer kliniskt beslutsfattande på ett människocentrerat och iterativt vis. Presenterade resultat kommer från fyra utforskande projekt som har utförts i samarbete med specialistläkare och har syftat till att utforska hur interaktiva system med ML komponenter kan utformas.

Presenterade bidrag inkluderar att identifiera praktiska och teoretiska utmaningar, föreslå lämpliga interaktionsstrategier för människa-AI samarbete inom medicinskt beslutsfattande samt att föreslå abstraktioner som beskriver egenskaper av maskininlärning och AI som är viktiga för designarbete. Specifikt inkluderas tre verk som behandlar att designa för system som lär sig av människor, två verk som beskriver designade beslutsstöd för specialistläkare samt ett verk som behandlar lättviktig prototypning och relevanta kompetenser för framtida AI-designers.

Acknowledgments

Looking back and pondering my experiences over these past five years, I am struck by emotions and memories that at times seems outlandish to my present self. Streams of changes and trials propelled me forever forward, to the extent that I am not quite the same person who started this research. I am genuinely grateful to all the people I encountered throughout this process, even if it was only for a brief chat as our paths crossed at an academic conference. However, this section calls for names, and names there shall be.

First, I would like to thank my main supervisor, Jonas Löwgren—your way of guiding, thinking and doing taught me about research in ways I could never learn from texts. I would also like to thank my other supervisors, Claes Lundström and Darren Treanor, for your advice and for creating opportunities for my research by connecting me with pathologists.

To all those who discussed my work and human-AI interaction with me, most notably Jesper Molin, Karin Stacke, Danwei Tran-Luciani, Apostolia Tsirikoglou and Qian Yang—thank you!

All my colleagues at Sectra contributed to a delightful working environment. A special thank you to the members of my research group, Daniel Forsberg, Erik Sjöblom, Olof Sandberg, Stefan Lindholm and Erik Gabrielsson.

This thesis would not have been possible without the enthusiasm and engagement some pathologists showed for my design experiments. I feel slightly ashamed calling them mine because they felt like our joint projects. Thank you, Gordan Maras, Karin Lindman and Arrigo Capitanio. Additionally, I am grateful to all the pathologists who allowed me to observe their work, tested my prototypes, or participated in my user studies. I apologize for not including your names.

Finally, these years have proven to be just as emotionally challenging as they have been intellectual. I owe my sanity to my family and friends. Grandmother Maj-Britt, you were there for me when there was no one else. I miss you. Stina and Lisen, your love changed my world.

For those of you still in this life, my deepest wish is that I can support you as you have me.

Martin Lindvall

Linköping, June 2021

List of Publications

This thesis is based on the following appended papers:

- Paper 1.** M. Lindvall, A. Sanner, F. Petré, K. Lindman, D. Treanor, C. Lundström, and J. Löwgren, *TissueWand, a Rapid Histopathology Annotation Tool*, Journal of Pathology Informatics, 11(27), 2020
- Paper 2.** M. Lindvall, C. Lundström, and J. Löwgren, *Rapid Assisted Visual Search: Supporting Digital Pathologists with Imperfect AI*, In Proceedings of the 26th ACM Intelligent User Interfaces Conference, 2021.
- Paper 3.** M. Lindvall, J. Molin, and J. Löwgren, *From Machine Learning to Machine Teaching: The Importance of UX*, Interactions, 25(6), 52-57, 2018.
- Paper 4.** D. Tran Luciani, M. Lindvall, and J. Löwgren, *Machine Learning as a Design Material: a Curated Collection of Exemplars for Visual Interaction*, DS 91: Proceedings of NordDesign 2018, 2018.
- Paper 5.** M. Lindvall and J. Molin, *Verification Staircase: a Design Strategy for Actionable Explanations*, In Proceedings of the Workshop on Explainable Smart Systems for Algorithmic Transparency in Emerging Technologies, 2020.
- Paper 6.** M. Lindvall and J. Molin, *Designing for the Long-Tail of Machine Learning*, In Workshop on Emerging Perspectives in Human-Centered Machine Learning at CHI '19, 2019.
- Paper 7.** C. B. Stadler, M. Lindvall, C. Lundström, A. Bodén, K. Lindman, J. Rose, D. Treanor, J. Blomma, K. Stacke, N. Pinchaud, M. Hedlund, F. Landgren, M. Woisetschläger, and D. Forsberg, *Proactive Construction of an Annotated Imaging Database for Artificial Intelligence Training*. Journal of Digital Imaging. 34, 105-115, 2021.

Other relevant publications authored or co-authored by Martin Lindvall:

Lindman, K., Rose, J. F., Lindvall, M., Lundström, C. Treanor, D. *Annotations, ontologies, and whole slide images – Development of an annotated ontology-driven whole slide image library of normal and abnormal human tissue*. Journal of Pathology Informatics, 10(1), 22, 2019.

Lindvall, M. Molin J. Löwgren, J. *The Importance of UX for Machine Teaching*. In 2018 AAAI Spring Symposium Series, Stanford, 2018.

Tsirikoglou, A. Stacke, K. Eilertsen, G. Lindvall, M. Unger, J. *A Study of Deep Learning Colon Cancer Detection in Limited Data Access Scenarios*. ICLR 2020 Workshop on AI for Overcoming Global Disparities in Cancer Care (AI4CC). 2020.

Published datasets co-authored by Martin Lindvall:

Jarkman, S. Lindvall, M. Hedlund, J. Treanor, D. Lundström, C. van der Laak, J. *Axillary lymph nodes in breast cancer cases*. 2019. doi:10.23698/aida/brln

Bodén, A. Rose, J. Lindvall, M. Bivik Stadler, C. *Breast data from the Visual Sweden project DROID*. 2019. doi:10.23698/aida/drbr

Lindman, K. Lindvall, M. Bivik Stadler, C. Lundström, C. Treanor, D. *Colon data from the Visual Sweden project DROID*. 2019. doi:10.23698/aida/drco

Lindman, K. Rose, J. Lindvall, M. Bivik Stadler, C. *Ovary data from the Visual Sweden project DROID*. 2019. doi:10.23698/aida/drov

Lindman, K. Rose, J. Lindvall, M. Bivik Stadler, C. *Skin data from the Visual Sweden project DROID*. 2019. doi:10.23698/aida/drsk

Maras, G. Lindvall, M. Lundström, C. *Regional lymph node metastasis in colon adenocarcinoma*. 2019. doi:10.23698/aida/lnco

Maras, G. Lindvall, M. Lundström, C. *Regional lymph node metastasis in colon adenocarcinoma, second collection series*. 2019. doi:10.23698/aida/lnco2

Personal contribution statement

- Paper 1.** Lindvall was the main contributor to the design and execution of the research study, the analysis of the data, the manuscript, ML models, algorithms and the interactive tool. Early design work was done by Sanner and Petre as part of their master thesis. Lindman and Treanor provided domain expertise and Lundström and Löwgren supervised the project and assisted in revisions of the manuscript.
- Paper 2.** Lindvall was the main contributor to the design, planning and execution of the research study, the analysis of the data, the manuscript, algorithms, ML models and the interactive tool. Lundström and Löwgren provided supervision and assisted in revising the manuscript.
- Paper 3.** Lindvall was the main contributor of the manuscript. Molin and Lindvall contributed equally to the design of the presented prototypes and identifying the underlying concepts. Molin and Löwgren assisted in revising the manuscript.
- Paper 4.** Lindvall and Luciani contributed equally to the concept of curated collections and its motivation. Luciani was the main author of the manuscript. Lindvall contributed with the exemplars and the suggested principles. Luciani made the work on applying the principles on aircraft automation. Lindvall and Löwgren assisted in revising the manuscript.
- Paper 5.** Molin came up with the idea of the verification staircase. Lindvall extended the initial concept and framed it as part of the explainable AI research topic. Lindvall and Molin equally contributed to the presented prototype. Lindvall was the main author of the manuscript. Molin assisted in revising the manuscript.
- Paper 6.** Lindvall was the main contributor to the preparation of the manuscript and the main contributor to the identified insights that are discussed. Molin assisted in revising the manuscript.
- Paper 7.** Lindvall was the main design and technological lead for the ovary AI prototype. Lindvall also assisted with case selection and annotation strategy across pathology-related projects. Lindvall assisted in drafting and revising the manuscript.

Contents

Abstract	iii
Acknowledgments	v
List of Publications	vi
I Comprehensive Summary	1
1 Introduction	3
1.1 Research Scope and Delimitations	5
1.2 Thesis Overview	5
2 Background	7
2.1 Machine Learning Basics	7
2.2 Digital Pathology	9
2.3 Artificial Intelligence in Pathology	12
3 Theoretical Framework	15
3.1 Automation and Human Control	15
3.2 Challenges Designing with ML	16
3.3 AI-assisted Decision-making	19
4 Research Approach	21
4.1 Constructive Design Research	21
4.2 Motivations	22
4.3 From Design Experiments to Generative Knowledge	23
4.4 Designing for Efficiency	25

5	Paper summary	27
6	Design Experiments	31
6.1	Overview	31
6.2	Exploring Proactive Training Data Collection (DROID)	33
6.3	AI-assisted Annotation (TW)	38
6.4	AI-assisted Visual Search (LGL)	42
6.5	AI-assisted Quantification (PDL1)	51
7	A Framework for Designing Human-Centred Machine Learning	59
7.1	The Importance of Thoughtful Action	59
7.2	Three Interconnected Activities	60
7.3	The Impact of Gathering Training Data	61
7.4	The Impact of Interaction Design	63
7.5	The Impact of Model Development	63
8	Conclusion and Discussion	65
8.1	Summary of Contributions	65
8.2	Why is Designing with ML Difficult?	66
8.3	Power to the People? Reflections on Interactive Machine Learning	69
8.4	The Role of Constructive Design Research	70
8.5	In Conclusion	71
	Bibliography	73
II	Appended papers	87
1	TissueWand, a Rapid Histopathology Annotation Tool.	89
2	Rapid Assisted Visual Search: Supporting Digital Pathologists with Im- perfect AI.	93
3	From Machine Learning to Machine Teaching: The Importance of UX.	97

4	Machine Learning as a Design Material: a Curated Collection of Exemplars for Visual Interaction.	101
5	Verification Staircase: a Design Strategy for Actionable Explanations.	105
6	Designing for the Long-Tail of Machine Learning.	109
7	Proactive Construction of an Annotated Imaging Database for Artificial Intelligence Training.	113

Part I

Comprehensive Summary

Chapter 1

Introduction

Interest in artificial intelligence (AI) has surged in recent years. Again, expectations on technological capabilities spark dreams of making digital things that benefit the human condition. These new things we envision might help us solve old challenges in new, more preferable ways or enable things that were not previously possible—creatures of our imaginations.

The vision for my PhD topic formed in the wake of the seminal results that deep neural networks had on the ImageNet challenge [33] of classifying natural, everyday photographs. The resulting predictive performance of this old-made-new machine learning (ML) method beat most old results. However, the aspect that attracted me was another. Compared to older methods, results were now achieved solely from labelled data, with less emphasis on advanced mathematical translations of problems into a form suitable for learning that characterised older ML approaches. It seemed to have the potential to be, in a sense, highly teachable without in-depth knowledge of image-related algorithms, filters, feature engineering or signal processing.

For the past decade, I have been working professionally within the medical informatics industry in what might best be described as a “prototyper” role. My professional work has typically combined interaction design, business analysis and programming for purposes ranging from innovation research to production systems.

Out of this background, I sought to explore the potential of the new machine learning technology for interaction design. The context of my PhD studies is in Sweden known as an *industrial PhD*: a position where employment remains within the industry, but one becomes associated with a university to pursue research questions and contribute to ongoing discourse with the academic community. This usually comes with an expectation that results have both academic and external relevance.

Associating with a particular external domain follows from this context. For my work, that domain has been medical image analysis for clinical pathology. In clinical pathology, specialised physicians analyse tissue samples from biopsies or surgeries in high magnification using a microscope or, within digital pathology, using a computer monitor. Based on visual analysis, decisions are made that affect patients’ treatment choices, choices that can affect the quality of life, or whether the wonders of life are allowed to go on at all.

The last decade's advancements in machine learning have led to a dramatic increase in AI capabilities. However, despite impressive technical advances and many successful research projects, machine algorithms for medical diagnostics are only minimally used in healthcare today. Traditional machine learning research has been criticised for having an overemphasis on presenting algorithms that do well in benchmarks of well known, static data sets. This has led to a lack of follow-through research, with impact and adoption in relevant user communities still unrealised [125]. Research on design considerations and relevant human-computer interaction issues has been identified as one potential area to close this gap [59]. There is a need for examples of what holistic success might mean when designing with ML; in real-world contexts, design practitioners need to navigate complex objectives that go beyond benchmark performance on fixed datasets [29].

Creating effective and efficient end-user interaction with machine learning introduces new challenges that necessitate a better understanding of human capabilities, behaviours and needs [6]. Although several research communities have made efforts in this area, including visualisation and human-computer interaction, such systems' design space is still relatively unexplored with many open research questions [57], [98]. For instance: how can user interfaces allow users to mitigate imperfect predictions? How can such user interfaces be sketched and prototyped, despite the technology requiring extensive data collection and engineering efforts to generate realistic scenarios of system behaviour?

Recent approaches that frame the use of automation in more human-centred ways include an "algorithm in the loop" view [49] and algorithmic experiences (AX) [3]. These perspectives can be contrasted with the unfortunate use of the term "human in the loop", in which the human is positioned as just another cog in the wheel of automation.

Putting humans first and automation second is especially relevant for the medical domain. The decisions made by medical specialists are of a different kind than, for instance, recommending music: the designed experience needs to be extra prudent, thoughtful and capture the unique and extensive expertise of medical professionals.

The aim of my thesis can be summarised in an explorative research question; *How might we design human-ML collaborative systems for efficient clinical decision making?*

The overall approach used for the work within this thesis can be characterised as explorative design or human-centred design. In such research through design, the research question is explored through experimentation with digital materials. Starting from a focus on user values, it involves creating and refining artefacts to simultaneously explore the space of what is possible and what is useful. The thesis describes this approach applied to the field of digital pathology in four design experiments, resulting in three highly functional interactive prototypes of ML systems. The prototypes target AI assistance for annotation, visual search for rare phenomena, and cell-level quantification.

Designerly exploration of possible futures involves engaging with materials, including sketches, prototypes and models. Since ML system behaviour is data- and technology-dependent, all experiments have considered training data collection and ML model development an integral part of the design effort. There are few prototyping methods established for designers to work with ML. As a consequence, I have worked directly with the technology in most design experiments.

While my design experiments aim to provide AI support for specific diagnostic tasks, the knowledge gained through making and experimenting with ML as a design material can be lifted from the particular to a more general level. Through reflection and analysis, a characterisation of the design space for human-machine teaching is articulated along with a design strategy for interactive verification and a framework addressing designerly competence when working with ML.

1.1 Research Scope and Delimitations

This work explores how one might design with ML to make systems where the resulting value is co-created between artefacts and humans in particular situations of use. The value of interest is pre-specified as achieving a particular goal: efficient clinical decision making. The focus on collaborative systems and clinical efficiency can be seen as two a-priori delimitations on the broader topic of designing with machine learning.

Further, the studies in this thesis are limited mainly to medical image analysis for clinical pathology. This domain choice is due to the practical use of resources: the time required for a practising researcher-designer to understand the problem domain and build a network of collaborators and informants is extensive. Although the limitation to pathology is pragmatic, it is arguably a fascinating research domain. Clinical pathology is currently a hotbed of ML and AI research in computer science. The field is transitioning from using microscopes to working digitally. Many pathologists have a positive outlook on the possibilities of technology to aid in their work, somewhat unsullied by previous AI winters and hard-to-use software.

Another delimitation is that in this work, machine learning and artificial intelligence are not considered in the sense of what they might be capable of in some future state. Instead, they are explored as what the technologies are capable of now. This limitation is due to a belief that even if no further technological progress happens, as designers, we are still many years and much research away from fully understanding how to wield this technology as a material to create richer experiences.

1.2 Thesis Overview

The thesis consists of two parts. Part I contains a comprehensive summary and background. Part II contains seven appended papers.

In Part I, the *background* chapter introduces machine learning, characterises pathologists' work, and summarises the current state of the art in applying ML for digital pathology. Subsequently, I review the research that contributes to the *theoretical framework* of my thesis. This includes challenges of automation, understanding machine learning as a design material, and AI-assisted decision making. Having presented both background and relevant theory, I motivate and describe my overall *research approach* in further detail. The following two chapters then describe the overall work. First, a *summary of appended papers* and their results are given in light of the presented theory. Then the exploratory design projects in which these papers were written are presented as *design*

experiments together with my subjective reflections related to the main research question. Chapter 7 synthesises my findings and suggests a *framework for iterative human-centred machine learning*, intended to advance a designerly understanding of how to design with machine learning. Finally, I *summarise and discuss my contributions* and their implications for advancing current theory related to designing human-ML collaborative systems for efficient clinical decision making.

The recommended reading order is to begin with Chapter 1-5 of Part I followed by the appended papers in Part II. Chapters 6 & 7 are intended for design researchers and readers seeking further characterisation of the conducted design experiments. Chapter 8 summarises and discusses the impact of contributions and is intended to be read last.

Chapter 2

Background

This chapter provides background on machine learning, pathologists' work, and recent advances in applying ML to support their work.

2.1 Machine Learning Basics

While this work is primarily situated within the field of design research, the descriptions of my design experiments in Chapter 6 rely on understanding the current capabilities and limitations of ML. This section provides a gentle overview of the relevant AI and ML nomenclature used throughout this thesis.

The term artificial intelligence (AI) is commonly used in two senses. First, it can refer to software systems that demonstrate capabilities previously associated with the human intellect. Second, it can denote a field of research centred on technologies, that are currently or was historically, inspired by striving towards specific human-like cognitive abilities. In the second sense, the key problems of the AI research field include knowledge representation, planning, reasoning, natural language processing and learning (ML).

This thesis focuses on ML as a concrete, tangible *technology* and explores how designers might use it. An alternative way to frame this focus is to assume that even if ML would remain largely as it is today, fruitful research on ML as a material for designers could potentially stretch decades. The term ML will be used over AI wherever possible to emphasise that the technology is in focus instead of a vision of future human-like capability. Some exceptions are made to avoid confusion when referring to established terminology used in contemporary discourse, e.g. *AI-assistance* or *human-AI interaction*.

Conventional software outputs are typically the result of logical rules applied to data. In contrast, machine learning entails learning outputs by example. A statistical structure, an *ML model*, is learned from data by observing many training examples. In *supervised learning*, the training inputs are the instances (X), such as images, and the desired outputs (Y), such as whether the pictures contain cancer or not. Additionally, one needs to choose a *learning algorithm* and a metric that tells whether the algorithm is doing a good job, a *loss function*.

Before training, input examples (X and Y) need to be transformed into mathematical representations. Historically, applying machine learning algorithms on images involved pre-defining complex transformations in a challenging engineering process known as *feature engineering*. Feature engineering is task-dependent and involves capturing the essential characteristics of the image to help the learning algorithm. It can involve computing representations of overall texture, finding edges and determining the roundness of objects.

The technique primarily used in this work is *deep learning* through *convolutional neural networks* (CNN). The success of deep learning is partly due to the fact that the raw RGB values of pixels can be used directly as inputs, and still outperform older approaches. Compared to older machine learning, this further shifts development away from manual programming. In the context of designing systems to support medical decision making, the implication is that domain experts now have an even more central role in selecting, curating and annotating training data.

Applying *deep learning* involves selecting a *network architecture*, which can be thought of as a structure that determines the sequence of representation-transformation steps (*layers*) that can be learned during training. In a typical *classifier* architecture, such as ResNet [50] or DenseNet [58], the representation just before the final output is sometimes used as an abstracted representation of the image. It is then called an *embedding* which can have several uses, such as the similarity-based *clustering* used in the design experiment on AI-assistance for visual quantification in section sec. 6.5. A summary of nomenclature for ML tasks is given in tbl. 2.1.

Table 2.1: Nomenclature used for common machine learning algorithms

ML Task		Description
Clustering	Finding data points that are similar to each other and grouping them.	
Classification	Determining the class of an instance, such as A, B or C.	
Detection	Determining the presence and possibly the spatial location of objects of interest	
Segmentation	Translating from pixels into objects or assigning a class to each pixel	
Recommendation	Suggesting items based on similarity	

The result of model training is an *ML model*. Given an unseen instance X, it will produce an output Y with the same structure as training samples. The result can be associated with uncertainty. For an example classifier determining whether X is of type A or B, the output could be $A : 0.4$, $B : 0.6$ indicating that B is the most likely type, but that A is also somewhat probable. This is referred to as a *confidence score*. While this number should not be equated with a statistical probability, the word *probability* is sometimes used. In the case of a two-class problem, for instance, cancer or not-cancer, the *confidence threshold* (e.g. $\text{cancer if } > 0.8$) is usually set by evaluating performance on data withheld from training. In the design experiment on AI-assisted visual search in sec. 6.4 and Paper 2, there is no fixed threshold; instead, probable regions are sorted by this number.

Figure 2.1: Machine learning workflow from Amershi et al. [5]. Two loopback points indicate iteration, where the process can go back to any previous stage.

While ML models can appear to behave unpredictably, they are strictly deterministic in the sense that the same input will always generate the same output. However, if new training data is added, this guarantee is lost. How well an ML model manages unseen data is referred to as its ability to *generalise*. Our ability to know this generalisation capability, and thus the model’s true performance, is limited because enumerating all potential instances that it will be applied on is infeasible. Furthermore, performance can deteriorate due to changes that a human might find hard to notice, such as changes in the relative intensity of colours and the equipment that produced the image.

Software *engineering processes* for constructing ML models usually involve determining requirements, training data collection, model training, evaluation and post-deployment monitoring. One example is the ML workflow described by Amershi et al. [5], which divides the process into nine stages with two feedback points at which the process might loop back to any previous stage, as seen in fig. 2.1.

Unless used solely for research, an *ML model* needs to receive inputs and produce some action based on its outputs. In this thesis, I call this larger unit in which the ML component is situated an *ML system*. The notion of ML system excludes the user, and I will sometimes refer to the system with users included as *human-machine ensemble* or *joint system*. While the touchpoints where a user interacts with an ML system are the focal points of *human-AI interaction design*, as I will later argue, the design effort needs to consider the entire *ML system* in order to achieve the desired user experience.

2.2 Digital Pathology

The design experiments in this work have been limited to routine cancer diagnosis in histopathology using digital pathology. This section provides the necessary background in terms of context, technology and working practices in which digital pathologists are situated.

Histopathology is the microscopic study of tissues with respect to the manifestation of diseases. *Digital pathology* involves reviewing tissue specimens using a computer monitor instead of a traditional microscope. Although digital pathology systems have been available for research and education for decades, they have only recently started to receive larger-scale adoption for routine diagnostics [118].

Our human bodies are composed of *cells*. When cells cease to respond to extracellular signals in a way that causes them to grow and divide excessively, it can give rise to a *tumour*. When a tumour spreads into surrounding tissue, it is considered *invasive* and likely to be *malignant*. Some benign tumours remain relatively stable, and only malignant tumours are considered to be *cancers*. Invasive tumour cells can sometimes travel through the blood or lymph vessels and form new tumours. When this happens, the secondary tumours are called *metastases*, and the cancer is named after the originating tissue, such as when breast cancer is found in lung tissue. Cancer diagnosis is multidisciplinary and can, besides pathologists, involve specialists such as oncologists, surgeons, and radiologists. The role of the histopathologist is to analyse tissue samples in order to characterise the cancer disease, if present, and communicate their findings through diagnostic reports or other means. Patients' tissues are usually collected through biopsies or surgery and sent to the pathology lab for processing and analysis. The reason behind collecting the tissue varies, from a suspected abnormality to characterisation of tumour and surrounding tissue post-surgery.

Before tissues can be microscopically reviewed, they are cut into pieces, chemically processed and embedded into paraffin blocks, sliced into micrometre thin sections and placed on glass slides. Cells are colourless and appear transparent. In order to distinguish structures, slides are stained with chemicals that highlight features of interest. One common staining is hematoxylin and eosin (H&E), as in fig. 2.2, which renders cell nuclei a purplish-blue colour and cytoplasm a brighter pink. More advanced stainings involve *immunohistochemistry* (IHC), whereby a visual marker is constructed to attach to specific molecules. Typically these are proteins involved in cancer pathways, such as the PD-L1 molecule mentioned in Paper 5 and sec. 6.5.

To digitise glass slides, they are batch-fed into a whole slide scanner that produces high-resolution *whole slide images* (WSI). A 15x20mm surgical specimen results in a digital image of 5 gigapixels when scanned with a resolution of 0.25 micrometer per pixel (~"40X"). Using lossy compression, one such image file is roughly 1.5 gigabytes. In a digital pathology system, pathologists navigate these images through zoom and pan operations [102] in a manner similar to how modern map services allow navigation of satellite imagery.

Relatively few works exist that provide a high-level description of histopathologists' diagnostic process, and most works are based on conventional microscopic review. Pohn et al. [97] suggest a linear model consisting of four phases based on the empirical work by Crowley et al. [31]. In the initial *orientation phase*, the goal is to get an overview of the case, followed by a *search phase* where abnormalities are identified. In the subsequent *reasoning phase* several hypotheses are developed and tested through iterative searches. Finally, in the *closure phase* the diagnosis is summarised. While this model is easy to understand, it is likely an oversimplification with the phases corresponding more to how the experimental protocol of Crowley et al. was constructed.

In the original protocol analysis by Crowley et al., identified operators were grouped into categories consisting of examination, exploration-explanation, interpretation, goal-setting and control processes. Some examples include *evaluate-own-ability* as a control process and *evaluate salience* as part of exploration-explanation [31].

Figure 2.2: Scanned whole slide image (WSI) of colon tissue stained with H&E. Only a small portion of the original image is shown, as indicated by the green rectangle on the thumbnail of the whole image in the top-right corner. From the DROID colon dataset [16].

In a study based on observation of routine pathologist work, Randell et al. [101] analysed activities performed by four gastrointestinal pathologists. Results indicated that several activities frequently overlapped or were rapidly interspersed, including viewing the glass slide, dictating findings, making notes on paper and referring to additional case information such as the clinical details and the original specimen. There were also individual differences, such as some only dictating at the end of a case, while others did so intermittently. Notably, visually reviewing slides was the dominant activity for only 57% of the average case time, indicating that support systems should also consider the other activities involved.

Studies that analyse navigation patterns indicate that individual differences and expertise levels also account for noticeable differences. Two strategies commonly mentioned are *dip-zooming*, zooming quickly in and out, and *cover panning*, systematically covering an area in high magnification [85], [89], [102].

In his thesis on diagnostic review with digital pathology, Molin [88] adapted the more general theory of sensemaking by Pirolli and Card [96] to the diagnostic process using some of the categories identified by Crowley et al. [31], as shown in fig. 2.3. In this model, the work is driven by two main loops, the foraging loop where information is collected and the sensemaking loop where findings are distilled in order to determine what to communicate through the report. This model does not dictate that hypothesis formation precedes evidence collection or vice versa, consistent with findings from Crowley et al. [31]. In addition, this is consistent with the observation that some pathologists use a text-based report template as the driver for their work, while others collect evidence first

Figure 2.3: Sensemaking loop of pathologists. From Molin [88], reproduced with permission. © Molin, 2016.

and then generate the report [24].

2.3 Artificial Intelligence in Pathology

In recent years, several studies using digital image analysis based on machine learning have demonstrated high agreement between pathologists and automatic results. This section summarises a few notable contributions. For a more comprehensive review, see [1].

The CAMELYON challenge and the associated datasets attracted numerous ML-researchers to the task of identifying tumour metastasis in lymph nodes [12], [76]. Most high-scoring approaches utilise machine learning using convolutional neural networks (CNN), sometimes adding random forests for the final classification [12]. In the evaluation, the automated approach is compared to a manual, unassisted review. One particularly strong approach was the LYNA [77] 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 [64], performing Gleason

grading for prostate biopsies [21] and identifying basal cell carcinomas [23].

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 [13], more variations in image quality and rare target phenomena. Recent guidelines on implementing digital image analysis for routine use emphasise extensive human monitoring, including that a pathologist always visually verify the result [19].

Given that most automated approaches will need to be interacted with to some extent, it becomes necessary to study the human-machine ensemble. Recent studies have investigated how AI-assistance impacts ensemble performance, mainly focusing on diagnostic or prognostic outcomes and how the workload changes in terms of time efficiency. 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 [113]. The assistive system consisted of a zoomable user interface visualising regions of interest using bounding boxes, colour-coded for prediction certainty. In a system for Gleason grading, AI assistance increased concordance for most observers [20]. In a study of a content-based retrieval system for pathologists, the system demonstrated the potential to resolve some difficult decisions by allowing users to refine results interactively [22].

AI-assistance studies have also highlighted issues with employing ML. In a study investigating AI-assisted detection of helicobacter pylori, positive cases were diagnosed faster but at the expense of accuracy and certainty for negative cases [136]. In a study of an image analysis system that allowed interactive corrections of predictions for assessing Ki67 proliferation score, using assistance improved results compared to manual eyeballing, but corrections occasionally worsened the results compared to the unaltered predictions of the system [17].

Based on an analysis of the design processes of three image analysis tools, Molin et al. [90] summarised user needs for automated tools in digital pathology. First, verification of an automatic result must be faster than performing the original task. Second, visualisations should be employed such that prediction failures are apparent. This includes not occluding the underlying phenomena. Third, pathologists' usually need to make a speed-accuracy tradeoff depending on the case at hand. Thus, automated tools should afford a dynamic pace. Finally, any output should be designed so that the pathologist can create the report that they need to communicate to clinicians.

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 benefits in experimental settings and systems that can improve patient outcomes in clinical use.

Chapter 3

Theoretical Framework

This thesis focuses on designing systems in which the value-in-use partly stems from components that produce predictions through ML models. The study of AI in the context of human-computer interaction (HCI) can be dated as far back as when the term was first conceived, but in recent years the focus has been renewed in a “third wave” of research [91].

In this section, I selectively highlight research that has informed my work and forms part of the scientific discourse I aim to contribute to. The section on *automation and human control* reviews research on automation to support human tasks and challenges associated with over-automation. In *challenges designing with machine learning*, findings on how ML affects designers and the design process are summarised. Finally, I describe findings from studies that deal with *AI-assisted decision making*, where machine predictions augment tasks rather than automate them.

3.1 Automation and Human Control

The introduction of intelligent behaviour into human-controlled processes is a long-standing research topic within human factors engineering. The research has typically focused on safety-critical environments like rescue operations, commercial aviation, and control rooms for nuclear power plants.

Several re-occurring problems have been identified and associated with the introduction of automation into these domains. These include *out of the loop* issues, where the human cannot resume manual control due to reskilling, *overreliance*, when the operator fails to identify a situation where control should be applied and *modal* issues that occur due to a mismatch of the mental model of the operator surrounding the systems responsibility, and the mode that the system is actually in [11], [48]. The emergence of such issues can generally not be predicted from simple system characterisations, such as participatory mode. For instance, the ability of an operator to detect failures is improved when exercising active control, as opposed to having a supervisory monitoring role. However, when the workload is high, the opposite is true [116].

One approach to designing man-machine systems is to construct conceptual models that can guide function allocation between humans and automation. The Levels of Automation (LOA) framework has historically been highly influential [95], [106]. In this model, automation can be described in degrees, starting from a low level with no automatic assistance, through intermediary levels where humans review and approve all actions, to the highest level where the computer both decides and acts without human approval. The LOA framework relies on the decomposition of tasks into functions that can be independently allocated. One common decomposition divides tasks into four stages: information acquisition, analysis, action selection, and execution. Function allocation decisions can then be guided by the capacities of the human and machine actors. Lists such as the Men-Are-Better-At / Machines-Are-Better-At (MABA-MABA)[42] explicate such capacities, e.g. machine memory excels at literal reproduction, whereas human memory is better at retrieving strategies and principles for versatile problem-solving.

The LOA and MABA-MABA approach has been criticised for neglecting the importance of human cognition in determining the emergent behaviour of the complex joint system. One issue is that the level of granularity chosen for any task decomposition seems somewhat arbitrary, considering how nuanced and complex human collaborative work can be [32]. Relying on the LOA framework could lead to an erroneous belief that high levels of control and high levels of automation are mutually exclusive [107]. Norman [93] suggested that the feeling of being in control could be positively affected by designing feedback such that system understanding and confidence are increased. Systems would need to build confidence progressively. In the beginning, the system should provide detailed accounts of automatic behaviour. With increased understanding, users might consider this level of information a nuisance but still want to request the information occasionally.

Another serious objection to the LOA framework is that replacing certain human functions with automation through substitution will change the joint system—exploiting an automation capability will create new human strengths and weaknesses [32]. Given that no models of human cognition can perfectly predict behaviour, this means that a priori function allocation is intractable, and when used, resulting systems should be considered only as hypotheses for further inquiry [52].

When designing systems, starting from conceptual models of existing working practices and allocating autonomous functions might limit the range of systems that can be imagined [105], [122], particularly when considering the capabilities of machine learning [92].

The above theories have treated automation generally, regardless of the specific technology that provides human-like capacities. The rest of this chapter deals more explicitly with ML systems' challenges and end-user interactions.

3.2 Challenges Designing with ML

Creating successful systems that involve machine learning components offer novel challenges compared to traditional software design. Motivated by an apparent lack of innovation and unmet potential of ML from UX practitioners and interaction design researchers,

Figure 3.1: Difficulties designing with AI mapped onto a user-centred design process. From Yang et al. [133], adapted with permission.

Dove et al. [35] surveyed challenges that this technology poses through a survey of UX practitioners. Several challenges were revealed. Practitioners reported that the technological capabilities of ML were difficult to translate to potential interaction design opportunities. In addition, design-led innovation through progressive refinement of prototypes was rare. In a survey with designers working with ML-powered products, Yang et al. [132] revealed similar results. They called for more work on imagining novel uses of ML, including exemplars that can guide and sensitise designers to AI possibilities and establishing abstractions that capture desirable qualities and capabilities of AI in use. Others have similarly argued that the technical complexity is the source of the difficulties observed in practice [53] and that user-centred design methods need to be modified to work for ML [39], [47].

Yang et al. [133] synthesised findings from literature reviews, interviews with designers and research-through design work in several domains. They proposed that the UX of AI is uniquely challenging to design due to two root causes: uncertainty surrounding AI capabilities and AI's output complexity, ranging from simple to adaptive complex [129], [133]. Yang et al. also identified issues resulting from these two causes, ranging from determining the feasibility of early concepts to anticipating edge-case behaviours in deployed systems, as depicted in fig. 3.1.

User-centred design methods typically rely on being able to explore the solution space by utilising sketches or prototypes that can elicit insights on the resulting user experience at an early stage. However, realistic capabilities might be hard to envision without an existing ML model, which might, in turn, require an unwieldy amount of training data [35]. One option might be to restrict design-led innovation to after model construction.

Figure 3.2: Simplified version of the ML Canvas from Zhou et al. [137]. Each of the six stages is associated with several issues related to the application scenario and the ML system. In the figure, two example issues have been selected for each stage. The wording of issues has been abbreviated compared to the original work.

However, such limitations are undesirable since human decisions affect ML system outcomes throughout the ML life cycle, from training data collection to model selection [46]. Thus, to design ML systems, technology-driven ML development processes and design methods need to be consolidated. One proposed consolidation is provided by Zhou et al. [137], who mapped UX issues of human-AI interaction to ML development stages in an ML life cycle canvas, as in Fig. 3.2. The purpose of the canvas is to guide designers to think about UX issues such as quality of predictions, transparency, user control and continuous learning while working iteratively through six stages consisting of data collection, model construction, model training, inference, performance evaluation and model update.

Prototyping the UX of ML systems using low-fidelity methods is a relatively unexplored area of research. Findings from interviews with practitioners working with natural language processing indicate that prototyping was rarely considered, and when employed, insights were only of limited value due to the compromises involved [55]. Some have suggested using Wizard-of-Oz experiments (WoZ), where a human simulates the output of the predictive model [18], [104]. However, there are indications that such attempts do not work in practice since ML models fail in ways that are unintuitive for humans [131].

Several toolkits have been proposed for rapidly prototyping ML systems. The ML model

in such systems might perform worse than a final model would, but allows observing UX consequences of prediction errors [114], [123]. Hybrid approaches involve employing toolkits and cloud-based ML services while actively modifying the prediction pipeline to produce certain shortcomings [131]. Alternatives to prototyping include facilitating explicit consideration of failure scenarios [55] and constructing data-aware scenarios [115].

3.3 AI-assisted Decision-making

Integrating ML components into user interfaces creates new opportunities for interaction design but is also associated with challenges. As generated predictions are associated with uncertainty, ML systems violate established usability principles such as producing consistent behaviours [8]. Predictions in the context of a decision support system can be helpful [134] but might also disrupt existing workflows and become a nuisance to its intended users [127]. Machine learning research has been criticised for overemphasising evaluation using abstract metrics on a limited number of fixed benchmark datasets, neglecting larger issues needed for real-world follow-through [125]. Recently, a research agenda has been proposed for examining *human-centred machine learning* (HCML) [46], [100]. The HCML perspective emphasises knowledge surrounding how humans interact with ML for constructing ML systems as well as end-user interactions. This section focuses on studies that deal with how ML components can affect decision-making in a mode of interaction where humans verify ML predictions.

Two concepts that have been used to theorise on how predictions presented in the loop of human decision making can affect joint system performance are *trust* and *explanations*. While trust can be viewed as consisting of many experiential factors, such as perceived confidence in resulting decisions, the study of trust is commonly operationalised as demonstrating proper reliance on ML predictions. Since predictions can be faulty, proper reliance involves overriding recommended decisions for misleading predictions, avoiding over-reliance. Humans are not oracles; thus, proper reliance also involves not overruling correct advice [81], [126], resulting in under-reliance. Properly *calibrated trust* should improve joint task performance compared to the human acting alone or the model deployed in an autonomous mode.

Green and Chen [49] investigated how presentation style affected a risk assessment task. They found that subtle changes in presentation affected the joint performance, despite the ML model's being the same for all treatments. For instance, first asking a user to complete the task manually and then presenting the automated score made users more likely to go with the model's prediction if given the option to do so. Prior knowledge on how the ML model is expected to perform can also affect reliance. Lu and Yin [81] showed that in the absence of information on how an ML model performs, users would learn reliance strategies based on observing how the ML model performs on cases for which they are themselves relatively certain. Lu and Yin also found that users would change their reliance strategy after receiving feedback on true ML model performance. This has implications for how ML model performance is communicated prior to deployment. Some have suggested best practices on ML model performance communication, such as

model cards [87].

There is a large and growing taxonomy surrounding various forms of explanations [37], [54], [91] that aim to counteract the black-box nature of deep learning-based ML models. The impact of explanatory techniques on calibrating trust will be considered next. Note that these techniques can also be used for other purposes not considered here, such as informing ML model engineering.

One readily available type of feedback is the prediction uncertainty of the ML model. The idea is that information of when the ML model is confident or uncertain can guide users' reliance on the prediction. However, experimental results indicate that factors such as the task, the performance of the ML model and the expertise of users affect joint performance outcomes. For instance, showing uncertainty was observed to increase joint performance in detecting fake hotel reviews [65], but no such gain could be observed on an income prediction task [135]. While some have suggested that showing prediction uncertainty merely increases the chances that humans agree with ML predictions [84], the factors that influence outcomes remain largely unknown. Similar results, indicating task-dependence for proper reliance, have been found for other explanatory techniques, such as showing feature contributions, similar examples and counterfactuals [126].

Studies of AI assistance and explanations for clinical-decision making indicate that the opacity of ML models is usually not an issue in itself, as the workings of the human body and many commonly used drugs can also be viewed as "black boxes" [103]. In a study with clinicians working in intensive care units, Tonekaboni et al. [119] found that explainability was primarily viewed as a way of justifying decisions to other human stakeholders. They also suggested that explanations that have no impact on the workflow should be avoided since they might induce cognitive overload and waste precious time [119].

Chapter 4

Research Approach

My research question, “**How might we design human-ML collaborative systems for efficient clinical decision making?**” is future-oriented and captures that there might be more than one solution to consider. The notion of human-ML collaborative systems alludes to qualities of the interaction through a metaphor of collaboration. The question is intentionally quite loosely defined, limiting the solution space rather than precisely defining the problem. This kind of research question is well suited to an exploratory approach. One such approach is *constructive design research* (CDR) [60], also known as *research through design* (RtD) within HCI [138].

In this chapter, I describe and motivate CDR as my chosen research approach.

4.1 Constructive Design Research

In general, the CDR approach is motivated by recognising that some knowledge is best gained through the *making* of things [60], [79]. This mode of knowledge accumulation is particularly useful in situations where issues are under-constrained—there is no clearly defined situation to be solved, no pre-defined set of methods to use, nor an exact criterion to when the goal is achieved.

According to Simon [109], *designing* is devising courses of action that can change existing situations into preferred ones. In this endeavour, the goal can be framed as an attempt to find an adequate rather than optimal solution. In CDR, designing is used as a method of inquiry [43]. The researcher is typically an experienced designer who employ processes from design practice.

In an engineering approach, a solution is derived from a rigorous analysis of a specific problem. In contrast, in a design-led inquiry, understanding of the problem space and the solution are co-evolved through making and evaluating artefacts [30]. Typically, initial problem framing is followed by conjecture of one or more tentative solutions, which are then evaluated. Insights gained can be used to re-frame the initial problem, including criteria of adequacy, followed by conjecture of new possible solutions.

What distinguishes research through design from professional design is that the researcher aims to make contributions that can be *generative*—that is, they can inform the

construction of future research and artefacts. What then constitutes generative *design knowledge*? Two possible candidates are the produced artefacts, or *instances*, and the advancement of *general theory* related to the problem space.

In the HCI field, novel user interfaces are commonly presented together with empirical user studies. In order to be generative, knowledge of one instance must transfer to the construction of other instances. User interface prototypes acting as design exemplars are sometimes viewed as embodying theory by the design choices made. They are, however, generally not considered to be design knowledge in themselves: much of the rich understanding gained through iteratively re-framing the problem space and exploring the solution space cannot be perceived by the “solution” alone.

Generalising from empirical observation of one or a few instances to general theory is complicated by the fact that a CDR approach usually starts from ill-defined problems and that the goal is to construct new futures rather than understanding existing situations. The CDR designer-researcher use descriptive theories, but knowledge contributions are typically centred on the relevance of these theories to the situation at hand, rather than actively refining or refuting them [43].

As an alternative to these two extremes, design researchers attempt to communicate *intermediate-level knowledge*—abstractions that lie somewhere between instances and general theory [56]. These combine empirical evidence and argumentative pieces in a form, where in contrast to objective truth, the reader has an active role in appropriating the material. Besides presenting artefacts and evaluations, the reader should be assisted in determining the scope and relevance of the work. This can be achieved in several ways, including providing a curated set of key insights and the design rationale that informed the process [78]. From an optimistic viewpoint, these kinds of contributions might over time combine in a programmatic way towards more generally applied knowledge [138].

Academic criteria of what constitutes valid design knowledge require that the contribution must be contestable, defensible and substantive [56]. *Contestable* means the contribution must be inventive and novel for the academic community. *Defensible* means the arguments or evidence can be accepted, i.e., that the argument is criticisable and is grounded empirically, analytically and theoretically. Finally, *substantive* means the contribution must be relevant, primarily internally within the academic community, but also with respect to external goals. For intermediate-level contributions produced through a CDR approach, relevance criteria also include the future *generativity* of the contribution.

4.2 Motivations

The material presented in the two previous chapters points towards broad unresolved questions underlying the main research question: how should ML be understood and utilised by designers? How can impressive computational results be translated into routine clinical practice?

Through machine learning, one can produce predictive models that perform well on tasks that were previously exclusively associated with human intelligence, such as recognising

and classifying various forms of cancerous cells and distinguishing them from non-cancerous cells. These models learn from data by example as partly stochastic processes attempt to optimise some pre-defined goal metric. This leads to results that are often correct but also occasionally and unexpectedly very wrong. Furthermore, as indicated by sec. 2.3 and sec. 3.3, performance at the joint system level might differ significantly from the ML model alone.

As an interactive material, machine learning poses novel challenges to an interaction designer:

- Predictions are produced with inherent uncertainty, and the uncertainty is not easily measured or communicated.
- Instead of being programmed by logic statements, it learns from input/output examples.
- The achievable performance or “intelligence” of an ML model for a given task is hard to estimate in advance.
- Performance assurances such as “97% accuracy” can be known for a pre-defined population of samples, but in use, the instances on which it will fail is hard to know and might not correspond to having the same properties that humans find difficult.

These lead to far-reaching consequences for existing interaction design practices, such as complicating iterative prototyping and a lack of established design patterns and interactive principles. These and other tensions (see sec. 3.2) have motivated some design researchers to frame machine learning as a new *design material* which designers need to learn to work with, posing questions such as to the appropriate technological knowledge a designer need to possess, applicable design processes, and design exemplars of successful designs.

As part of practice, designers apply conventions, rules-of-thumb, in their design processes. These are learned by experience or based on formalised knowledge, such as common visual and interactive *idioms* of windows-and-buttons user interfaces and mitigations to commonly encountered usability problems [28]. We might denote situations where there exists comparatively fewer of these structures and idiomatic elements in place as *non-idiomatic design situations* [120]. Given the accounts of how design practitioners struggle to leverage ML capabilities for innovative interactions [35], [128], the design of ML systems is arguably one such non-idiomatic context.

This further motivates the chosen CDR approach in that it is expected to produce new exemplars and intermediate-level contributions that can aid the formation of new conventions and idioms.

4.3 From Design Experiments to Generative Knowledge

My work is grounded in a series of projects, *design experiments*, that explore aspects of designing for machine learning in medical image analysis, starting from training data collection and working towards more efficient decision making. The projects are all oriented toward building prototypes that are functional enough for assessing the value criteria for the specific project.

The scope of this work, from which part of the design space to explore to project-specific research questions, was not known from the on-set. In line with the chosen approach, the direction grew from re-evaluation as the work progressed. What has been common is that in each design experiment, I use my experience as a code-fluent designer and attempt to work with machine learning to make things that produce value for the intended end-users. In this process, I combine knowledge gained from other researchers working with similar research questions with my own insights and reflections. As I work, I am not only iterating upon the project at hand but also on how to frame the work and knowledge gained such that it might be communicated to other design researchers.

The interweaving of evaluating systems empirically and reflecting on my experiences is visible in the appended papers of this thesis (see Chapter 5 for a summary). Some papers contain empirical evaluations of the final prototype of a design experiment. Others are argumentative pieces where I attempt to communicate insights in an appropriate form. The latter kind of contribution is additionally given more space within this thesis. Extended descriptions of the conducted design experiments are given in Chapter 6, providing a rich framing of personal insights. In Chapter 7, a tentative framework for human-centred machine learning is presented together with examples. The framework is intended to provide some much-needed structure to the discussion on the role of design for building ML systems.

Methods employed in the design experiments are mostly techniques that exist under the umbrella of user-centred design (UCD) or user experience (UX). The techniques include iterative prototyping, involving end-users in the design process, and user studies. In addition, design diaries were kept for analysing the progression of subjective understanding. In contrast with the dominant emphasis on early low-fidelity prototyping, the fidelity of prototypes moved quite rapidly into the realm of runnable interactive systems. This was due to a lack of good approximations and suitable methods for simulating realistic behaviour of algorithmic outputs—essential nuances of the interaction lie in the experience unfolding over time as an interplay between the human and the ML system. With regular “deterministic” windows-and-buttons style user interfaces, one can expect both the user and the designer to be able to extrapolate how a system would behave from sketches. Maybe due to lack of experience or stemming from something inherent in ML as a design material, this shared extrapolation was harder to achieve with systems that contain predictive and complex outputs.

I also came to build the ML models mostly myself. My goal at the on-set of my studies was to find collaborations with computer scientists for this part. However, it soon became apparent that treating both training data and optimisation goals as flexible elements of the design effort was necessary for rapid prototyping. Additionally, the desirable properties of an ML model in use, including performance, were hard to specify in advance. This created communication challenges when the collaborating researchers had ambitions of their own to produce novel results within the field of machine learning, where the perspective of value-in-use is quite foreign [86].

This working process leads to two kinds of knowledge contributions. First, the developed ML systems can be viewed as methods for the application domain or as design exemplars to designers working with ML. Secondly, insights across projects can be generalised

into forms of intermediate-level knowledge, such as describing the design space of these systems, informing design methods, or identifying properties of ML important for design.

4.4 Designing for Efficiency

The criterion for research to be generative for future research achieves the scientific goal of establishing a progression of knowledge. While this is appealing in principle, it has problems in practical use. By itself, it is both internal to the research field and neutral, not stating in what sense the progression is desirable. In addition, it is hard bordering on impossible to assay generativity at the time of publication. Therefore, some introduce a criterion of external relevance to CDR, requiring that the preferred state the design attempts to achieve is clearly stated, and support is given as to how the particular artefact has reached or progressed towards that state [138].

I have chosen to emphasise the criterion of external relevance for the design exemplars generated by my research. This follows in part from being situated both in industry and academia and from the apparent gulf between impressive experimental results of machine learning on predictive tasks in clinical pathology and the lack of their use in daily clinical practice.

There are several ways in which an ML system can be considered valuable for supporting decision-making. The focus on *efficient decision making* follows from adopting a value-in-use perspective on which desiderata to design for. In this perspective, the goal of design is creating something in which the desired values are co-created between the users and the artefact in the context of use [10], [26]. While the primary focus has been on efficiency, the traditional *usability* criteria of *effectiveness* and *satisfaction* have also been considered, along with more loosely defined aspects of the overall user experience.

In the context of expert medical diagnosis, positive values can often be related to impacting patient outcomes in positive ways. Since resources are limited, efficiency is often a priority since spending less time on one task allows more time for other patients or on making better decisions. Choosing efficiency before quality is mainly pragmatic; ML cannot easily become better than human teachers, and additionally, healthcare is usually cost-averse.

The choice of emphasising external relevance also follows from an ambition to be relevant to practising designers working with machine learning; requiring external relevance forces a grounding to the actual properties of machine learning as it exists today, not treating the technology as something capable of working magic wonders in the future, but instead as a tangible material, whose properties we need to discover.

Chapter 5

Paper summary

This thesis contains seven appended papers (P1-7). In this chapter, the appended papers are summarised based on the theory and background presented in the previous chapters.

Two papers (P1, P2) describe and evaluate the final prototype from their respective design experiment, TissueWand (also see sec. 6.3) and Rapid Assisted Visual Search (also see sec. 6.4). The other five papers aim at communicating abstracted knowledge gained across design experiments. Two papers (P3, P4) characterise part of the design space of designing for ML and argue for new ways to construct generative knowledge. One paper (P5) provides an intermediate-level abstraction of the central design strategy employed in the design experiment described in sec. 6.5. Paper 6 is an argumentative piece that aims to frame how ML and AI can be understood by designers and identifies technological knowledge that can aid design efforts. Finally, one paper (P7) suggests principles for successful collaboration in cross-disciplinary teams when collecting training data for machine learning.

Paper 1 (P1): TissueWand, a Rapid Histopathology Annotation Tool [75]. Machine learning creates predictive models from training data. Thus, examples are a necessity. Besides being a hurdle to technology-centred researchers, the probabilistic nature of models create barriers for designers relying on fail-early fail-fast prototyping since predictive models cannot be rapidly prototyped due to costly data dependence. In the associated exploratory work, we identified the effort of eye-hand coordination and the time required to perform detailed delineations as the primary hurdle for expert annotators. While other existing tools can aid annotation interactively, they have typically chosen to promote efficiency over accuracy or rely on pre-existing predictive models. High accuracy is typically needed for subsequent model training. The design experiment aimed to investigate how we might design an interaction that utilised unsupervised machine learning to speed up the annotation work without losing accuracy. Through iteratively experimenting with both algorithms and user interfaces, we constructed a tool that focuses on rapid real-time feedback. The primary experiential quality of TissueWand is tight coupling [2], where the tool is experienced as an extension, minimising friction between eye, hand and intent. We identify this style of fine-grained dynamic coactivity as an exciting venue for interacting with automation, as opposed to coarse-grained turn-taking.

Paper 2 (P2): Rapid Assisted Visual Search: Supporting Digital Pathologists with Imperfect AI [70]. Designing human-AI interaction is challenging for safety-critical applications such as clinical decision supports. A successful user interface requires the navigation of AI strengths and imperfections while at the same time fulfilling requirements on efficiency, quality, explainability, and trust. This paper presents an efficient and effective ML system that is the result of human-centred design work. An evaluation with six pathologists demonstrates that the suggested design can 1) reduce task time with maintained quality, 2) build user trust progressively, and 3) learn and improve from use. In the paper, design choices are related to the overall user experience through interviews. Implications of this work are discussed with respect to establishing trust, the role of explanations, learning from use, and devising collaboration strategies.

Paper 3 (P3): From Machine Learning to Machine Teaching: The Importance of UX [73]. Systems that can learn over time challenges designers to think beyond individual product releases and instead consider the broader lifespan of systems. This paper draws upon my design experiments to characterise and exemplify how designers can navigate this new design space when designing *teaching* aspects of learning-capable systems. In brief, we suggest bootstrapping when no training data is available by creating domain-specific tools, like TissueWand, or by creating valuable tools without ML that allow collecting the needed training data. After achieving sufficiently intelligent tools, we suggest focusing on the usability of verifications and corrections to allow ML systems to be used productively. At this later stage, the designer also needs to take care to provide the proper incentive for providing feedback without deteriorating the overall user experience. Finally, we hypothesise that ML systems will need to be re-designed as the capability of the underlying model improves.

Paper 4 (P4): Machine Learning as a Design Material: A Curated Collection of Exemplars for Visual Design [121]. When viewed as a design material, machine learning is arguably more unpredictable and emergent than traditional ones. These properties motivate approaches such as CDR for researching the design of ML systems. However, insights gained from making artefacts in design research need to be abstracted from the particular to a higher level to form more generally applicable knowledge contributions. This paper suggests curated collections, a variation of annotated portfolios [44], [78], as a way to abstract and communicate intermediate-level knowledge. A curated collection presents thoughtfully selected and inter-related exemplars, articulating their salient traits in order to inform future designs. One curated collection addressing the fine-grained details of interaction with ML predictions is provided as an example, highlighting traits including the preservation of visual context, rapid real-time refinement, leaving traces, and applying judicious automation.

Paper 5 (P5): Verification Staircase: A Design Strategy for Actionable Explanations [72]. What if the trust in the output of a predictive model could be acted upon in richer ways than a simple binary decision of accept or reject? Designing AI assistance for medical specialists entails supporting a complex but safety-critical decision process. Commonly, decisions in this domain can be decomposed into a combination of multiple smaller decisions. In this paper, we present Verification Staircase—a design strategy that can be used for such scenarios. The verification staircase is when multiple interactive

assistive tools are combined to allow for a nuanced amount of automation to aid the user. This can support a wide range of prediction quality scenarios, spanning from unproblematic minor mistakes to misleading major failures. By presenting the information in a hierarchical way, the user can learn how underlying predictions are connected to overall case predictions, and over time, calibrate their trust so that they can choose the appropriate level of automatic support.

Paper 6 (P6): Designing for the Long Tail of Machine Learning [71]. In this position piece, I draw upon experiences from my design experiments to first re-frame the typical human-AI interaction design process to include data collection and model development explicitly. This broader framing creates both opportunities and challenges for designers and product teams. Specifically, user interface improvements must now compete with improvements possible through data collection and model development work, and vice versa. While there is no single answer on how to navigate such situations, I suggest that basic knowledge of how ML model capability scales with data is part of the ML competence needed to enable designers to direct development efforts productively. The paper provides an early version of the framework presented in Chapter 7.

Paper 7 (P7): Proactive construction of an annotated imaging database for artificial intelligence training [112]. The availability of training data is crucial to developing ML systems for oncology imaging. This paper describes a large proactive data collection project that resulted in seven datasets originating from pathology and radiology. Besides describing the datasets, the project and experiences are presented as an exploration of best practices. The main contribution is generic lessons learned regarding successfully organising efforts to construct medical imaging databases for ML training, summarised as eight guiding principles covering team, process, and execution aspects.

Chapter 6

Design Experiments

The research contributions presented in the appended papers and summarised in the previous chapter mostly omit detailed descriptions of the design processes used. However, describing the design choices made in the context of the chronological unfolding of projects presents an opportunity for highlighting salient aspects relevant to how we might design with machine learning.

This chapter provides an abbreviated account of the four design experiments that have been the basis of the resulting papers. After a brief overview, each design experiment is described in terms of the design situation and the design process. Each section is followed by a few insights related to the main research question of this thesis. These accounts and insights have been produced by analysing the design diaries, sketches, intermediate prototypes, observation notes and interview transcripts collected during each project.

The process accounts and intermediate insights presented will lead to the next chapter, where they are used as empirical grounding and synthesised into a framework for guiding action when designing with machine learning.

6.1 Overview

The work consists of four projects that all were started at roughly the same time, as depicted in tbl. 6.1. In the projects, I have, with one exception, been the main contributor in implementing machine learning algorithms, evaluating models, sketching, coding and conducting user studies on interactive prototypes.

All projects have involved clinical experts in a relation closer to collaborators than primary informants.

Table 6.1: Projects and their major milestones. Field study: studying the intended situation of use at the clinic. Model Dev: whether an idiomatic (traditional) ML prototype was built as a starting point for iteration. IxD: iterative interaction design. User study: evaluating a high-fidelity prototype in use. Paper: whether the results have been published and if the results are appended papers of this thesis.

Project	Start date	Field Study	Model Dev.	IxD	User Study	Paper
DROID	2017-08	Y	Y	-	-	P7
TW	2017-08	Y	Y	Y	Y	P1,P3,P4
LGL	2017-11	Y	Y	Y	Y	P2,P4
PDL1	2018-01	Prior work	Y	Y	-	P5,P3

6.2 Exploring Proactive Training Data Collection (DROID)

6.2.1 Design situation

The aim of the Diagnostic Reference Oncology Imaging Database (DROID) project was to support future AI training through constructing datasets of medical images with associated annotations. The main intent was to be proactive in the sense of allowing the datasets to be generally useful for a wider range of oncology tasks. As the future usefulness would be hard to assess, the work also included defining a few tasks on which machine learning models could be developed end-to-end. The project had two working groups, focusing on radiology and pathology, respectively. Each group consisted of representatives from industry, academia and clinical practice.

My participation in the project took place between January 2017 through January 2019. Most participants, including me, lacked experience of similar end-to-end ML projects. While I had worked with ML using pre-existing datasets, the wide scope and the cross-discipline collaboration was new. I was situated in the pathology group and had dual roles as, on the one hand, an advisor on how to match clinical relevance with technological possibilities across subprojects, and later, to lead the development of one example end-to-end project, namely tumour type classification of ovary tissue.

The ovary prototype targeted the prediction of whether additional immunohistochemistry (IHC) staining is needed for clinical diagnosis. High-grade serous carcinoma (HGSC) is the most common type of cancer originating from the ovary/tuba, characterised by low differentiated cells with a high mitosis rate. The standard H&E stain is, however, insufficient to set the HGSC diagnosis as other carcinoma types can have a similar morphology. If any of these cancer types are found, the clinical process is to order additional tissue sections stained by certain IHC markers. The period female patients spend waiting for their cancer diagnosis can be emotionally taxing, with uncertainty causing anxiety and worry. Predicting the need for the additional staining and triggering it automatically, while not reducing pathologist work per-se, could lead to shorter turnaround times and has been suggested in prior work [41].

The type of system that was imagined can be described as highly agentive; it would do background work, immediately ordering chemical post-processing of tissue that typically has a lead time of a few days. Highly agentive systems should not be considered free from human interaction, as human control will likely involve tuning and monitoring [92]. I was curious as to how tuning and monitoring interactions might be realised within the pathology domain. Which need would there be to manually include or exclude cases for processing? Where at the clinic would this be done, by whom? How would pathologists react to cases where IHC was unnecessarily ordered? When missing, what kind of feedback in order to improve the prediction would they be willing to engage with?

However, these issues proved hard to engage without a realistic predictive model of sufficient quality.

6.2.2 Process

The phases of the project were matchmaking, annotation strategy development and execution, and model development.

At the onset of the project, we held matchmaking workshops [14] to evaluate which medical areas within oncology to target. The criteria unfolded in discussions and came to revolve around clinical relevance in terms of demand for AI assistance, the feasibility of a successful ML system, and whether the project time span would allow creating sufficient training data. A more detailed description is given in Paper 7. In pathology, four areas were selected, breast, ovary, skin, and colon.

The work for the ovary project proceeded through iteration of data collection, model training, and refinement of sketches surrounding the value proposition. The interactive aspects, imagined as monitoring and tuning tools, were difficult to realistically refine without a suitable predictive model situated in practice. While the performance of the produced model was promising, we decided against proceeding due to the high cost of realistic prospective experiments and the perceived importance of such realistic clinical integration for generating relevant insights.

Besides the ovary-specific project, I involved myself in the data capture and annotation efforts of the pathology group for all four areas. This was done because it turned out to be the bottleneck of subprojects, due in part to the time-consuming nature of annotation and the scarce availability of expert annotators. In addition, deciding what and how to annotate in order to be effective proved to be challenging. My engagement included observing annotation work and discussing annotation strategies in structured as well as informal settings.

6.2.3 Resulting artefacts

My engagements resulted in non-interactive artefacts, including one predictive model and four datasets.

The model produced was a predictive model for ovary subtyping, deciding whether to order an IHC or not based on the H&E image. It could be useful to inform the design of a monitoring and tuning system in future works. The prediction goal was to place high-grade serous (HGSC), endometrioid (EN), clear-cell (CC) and metastatic serous (MS) into an IHC-order classification, and place mucinous (MC), metastatic other (MO), low-grade serous (LGSC), mucinous borderline (MB), serous borderline (SB) and normal tissue (N) into a non-order classification. We achieved a rate of 11% false orders (4 out of 35) to a correct IHC-order rate of 60% (17 of 27).

The datasets produced have been made available through the AIDA datasets register [9], including the breast [15], colon [16], skin [69], and ovary datasets [68]. An example image from the skin dataset is shown with annotations in fig. 6.1.

Figure 6.1: Example region of an annotated whole slide image from the DROID skin dataset [69]

6.2.4 Main insights

Annotation considerations benefit from being iteratively discovered. Early model outputs can act as a boundary object, facilitating discussion on navigating speed-effect tradeoffs. Experienced pathologists had a tacit understanding of their classification processes and were not used to translating their diagnostic assessments to individual pixels. This issue was identified early on, but we still struggled to communicate about the issue effectively. Only when we had created initial ML models and observed problematic predictions could we judge the relative value of different annotation strategies. Discussions surrounding the model output provided a way to bridge the gap between the machine-centred, pixel-oriented characterisation and the intuitive processes of pathologists.

For instance, for the ovary project, we debated how to treat stroma. The pathologists remarked that it was an open question whether there is prognostic value in stroma and that cancer-related stroma looks somewhat different than normal stroma. Initially, we included stroma in the same annotation as the tumour type, as explicitly delineating stroma would have increased the annotation effort. After we had trained the initial ML model, one pathologist remarked: *“A thought that struck me was how I look at the slides. First, I start with whether I see any growth pattern, and after that, I look at the cells. In case one, the growing pattern of tumour cells is not typical for high-grade serous carcinoma, but in higher magnification, the cells look like high-grade serous carcinoma (at least low differentiated tumour cells), making it hard to decide if it is typical or not.”*. Initially, the annotating pathologist had annotated for the overall growth pattern, but our prediction model struggled to go beyond cell characteristics. There is no clear answer to whether changing the annotation strategy was the right way forward; adjusting the model to combine growth patterns and cell characteristics was also an option.

In a second step, we used TissueWand (Paper 1) to annotate all stroma rapidly. This was performed in roughly two days. The model improved considerably. This example illustrates how use-case, model development considerations and annotation strategy interweave. A reasonable question is whether we should have annotated stroma from the start to save time. In our case, annotating this post hoc did not impose a significant time penalty, and given that we did not know the information content of stroma, it seems more sound to prepare for this kind of iterative drift of the annotation strategy to occur.

In contrast, the skin and colon subprojects lacked an associated ML prototype effort. These projects required numerous more hours of annotation and the pathologists struggled to establish criteria for what phenomena to annotate and at which level of detail, to be useful for future, unspecified, ML prototypes. While it is impossible to determine the future value of the skin and colon datasets compared to the ovary dataset, at the time of writing, the ovary dataset has been used and cited in derivative works while the former has not. Even if some of the many ontological classes produced in the unguided projects prove valuable, the chances are that overall, the hours spent producing detailed annotations will be wasted since the annotations are not specific enough for any single AI task.

Anticipating and imagining the role of AI, without real AI, is challenging The main research ambition with the ovary prototype had been to explore what kind of

interactions would be effective and valuable in an agentic scenario, with the model ordering immunostains in the background. However, to anticipate needs realistically, we felt we needed to understand what kinds of errors could occur, which roles in the pathology department would be best suited to intervene and to what extent they could provide positive value through intervention. Understanding current working practices would not be enough; we would need to reimagine parts of the workflow. In order to engage stakeholders in this process, the benefits and drawbacks would need to be realistic. Creating a realistic working situation at the clinic was left as an interesting but experimentally challenging future avenue of research.

An important insight was that, in order to not overinvest in a novel concept that might be unwieldy to realise in practice, design should start from a realistic understanding of a predictive model trained for the given problem, including its shortcomings. The initial model might be rough, but insights from it form a technological perspective on which one can pivot towards a human-centred perspective.

6.2.5 Concluding remarks

The DROID project yielded insights that provided starting points for subsequent design experiments. It also exemplified how ML projects might be procedurally carried out to progress from initial matchmaking to predictive models. However, it remained unclear how consideration of a specific user interaction might change this working process and the factors that would need to be considered.

What was clear, however, was that training data is intimately connected to the design effort. Similar in spirit to the importance of UX to *building the right thing*, so is the importance of *learning the right thing* [130]. Two aspects of the training data are crucial to success; the form of the training data embeds assumptions about what kind of task it is that is solved, and secondly, the efficiency of the tools available to generate that data determines in large part the cost, and feasibility, of developing the solution. Thus, the training data can act as *foci* for the crucial design decisions at the early stages of development.

My experiences from the project, from thoughts on annotation, concept drift and the need for an iterative methodology was partly rendered into Paper 7.

Our work with constructing taxonomies for annotations is described in another work[67], not included in this thesis.

6.3 AI-assisted Annotation (TW)

6.3.1 Design situation

Machine learning entails learning by example, and in most cases, more data means more robust performance. Lack of data has been identified as a severe challenge for machine learning. The time and expertise needed to produce training data usually make it expensive and thus an attractive resource for researchers and industry alike.

Many aspects make annotation challenging: case selection and extraction, constructing a taxonomy and an annotation strategy that captures the correct information for the envisioned prediction task, and the arduous process of labelling pixels. In this project, we set out to support the latter, with a vision to use intelligence to build intelligence: to design an interactive, semi-intelligent tool that could be used for faster annotation of digital pathology whole slide images. The project was a collaboration with a group of researching pathologists who were manually annotating a collection of skin, colon, ovary and breast tissues in an early phase of the DROID project, as mentioned in the previous section.

Aiding a human to segment an image into constituent parts is not a new idea, and several tools have been proposed by researchers or made available as professional software. However, the end goal of using the delineated segments for training ML is different. The use of annotations for ML demands having high quality, low noise annotations. In most prior work, speed has been prioritised at the expense of quality. For example, if the end goal is to change the background for a photo, one might be satisfied with achieving 90% accuracy in a short amount of time. However, when producing training data for ML, we desire at least as high accuracy as the manual approach can produce. In addition, most previous tools have not been developed for gigapixel size images and the texture-rich context of tissue slides. A more extensive review of related work is given in Paper 1.

The research question driving this exploratory design project was: How might we design an annotation tool for whole slide images that produces high-quality delineations while being faster and less exhausting than polygon-based drawing tools?

6.3.2 Process

Overall, we adopted a design-led approach where ideas on technical properties and interactive behaviours were sketched, refined, and assessed in an iterative process.

The design effort started with observing the manual annotation task and reviewing prior work. Then, the process consisted of iterating upon algorithms, visualisations, interactive behaviours and the desired goal throughout as one whole. Due to the uncertainty on how algorithmic and interactive elements would fuse, high fidelity prototyping was used from an early point.

We identified the effort required for eye-hand coordination as the primary hurdle for expert annotators. Typically, pathologists could quickly point to the screen and instantly

Figure 6.2: Rejected prototype. Drawing a scribble (left) and receiving the results (right) in the context of a colon case.

remark ‘this is tumour’, but translating those almost-instant characterisations to detailed polygons could take from 10 to 300 minutes per image.

We evaluated a few available annotation tools with a predictive component. They typically utilised a workflow consisting of: annotate, predict, then refine and repeat. For example, Ilastik [110] provides predictions on the entire image after each refinement step. After receiving initial predictions, subsequent steps can be characterised as a task of verification and correction. For verification, the user must check the output to find errors and then use some correction interaction to refine results. When using gigapixel images, the verification task can become overwhelming due to potentially having to scan the entire image after each refinement. To mitigate this, we initially sought a workflow that limited the spatial scope of produced predictions.

An early prototype used an idiomatic, scribble-based input-output method but limited in spatial extent and adapted to histopathology. However, it became apparent that the experience of using this tool was unsatisfactory. One problematic behaviour observed was related to scribbling and waiting for results, as depicted in fig. 6.2. The user would typically scribble, receive unsatisfactory results, make some corrections, possibly overextending over the required bounds, erasing and starting over. This process could continue for quite some time, and sometimes the desired state would not be reached, inducing frustration.

The intent became to reduce the negative impact of unanticipated predictions errors. In order to achieve this, first, we sought to speed up the feedback loop to near real-time. By observing errors quickly, users might fail fast and limit the severeness of errors. In addition, we wanted to provide a mechanism to control the automation level so that the level could be reduced as a mitigation strategy when predictions were poor and increased to exploit predictions when possible. Taken together, this might allow users to develop an understanding of the underlying algorithm that is actionable.

Figure 6.3: Rapid feedback in the revised version of the tool. Each fat round dot represents a point at which the resulting segmentation was updated. The rapid mouse movements indicate that the feedback cycle is fast.

After the design phase, we conducted a small user study with a between-subject design to assess the prototype.

6.3.3 Resulting artefacts

Our final prototype, TissueWand, was a superpixel-based interactive tool that produces segmentations spreading from the user's cursor with a spatial extent controlled by applied stylus pressure, as in fig. 6.3. The feedback was real-time, with fast response times (~40 ms), creating a feeling of a tightly coupled human-machine system. A more detailed description is available in Paper 1 [75].

In a small user study, TissueWand was preferred to drawing polygons and, on average, needed only 57% of the time required for manual annotation.

6.3.4 Main insights

Rapid fine-grained interaction can mitigate AI unpredictability. Machine learning outputs can be unpredictable. Our strategy in TissueWand was to provide rapid feedback over a limited spatial range, such that prediction, verification and correction become so close in terms of time that they might not be experienced as separate. This fine-grained flow can be contrasted with framing human-automation interaction as coarse-grained turn-taking.

Design of machine-learning-based annotation aids should consider the timeline of predictive capability. The predictive capability of ML-derived models can increase with available training data. From this follows that an interactive annotation tool designed for the beginning of such projects, when the predictive capability is poor, might be relatively inefficient in later stages when fewer errors are produced overall.

The predictive component in TissueWand uses unsupervised clustering on pre-determined features. During the design process, we experimented with introducing supervised learning as the number of annotations increased. However, we identified two significant challenges with this approach. First, how should learned similarity be integrated with

the pre-computed similarity? We could imagine that learned similarity would allow spreading further and more accurately, but when during the annotation effort should this be exploited? If done before enough representative samples of all classes had been annotated, it might affect the experience negatively. Second, using the rapid fine-grained interaction of TissueWand might be under-utilising the model at this point. For instance, it might be more important to use approaches such as active learning or model debugging to identify informative samples rather than merely annotating more pixels.

This teaching aspect of the ML design space might be envisioned as a dimension of predictive capability over time, with differing design considerations along the scale. On such a dimension, TissueWand is a bootstrap tool. Characterisation of design considerations at further points might be more fruitfully explored through projects starting from existing predictive models rather than naively improving TissueWand. For instance, tools designed for highly capable ML models might not be annotation tools but rather correction mechanisms built into deployed products.

6.3.5 Concluding remarks

The exploration produced TissueWand and yielded insights into the role of training data for producing ML models. It also demonstrated the need for early prototypes to explore both interface- and computational-level behaviours.

The design and evaluation of TissueWand are described in more detail in Paper 1, and characterisation of the larger design space of designing for data gathering is described in Paper 3.

6.4 AI-assisted Visual Search (LGL)

6.4.1 Design situation

In this design experiment, an ML-powered system was developed for aiding a diagnostic task routinely performed by pathologists: assessing lymph nodes for tumour metastasis. The problem was selected because it is an active topic for computer scientists and is considered to have strong results algorithmically (see sec. 2.3). By exploring the design of decision support for a problem with strong automatic performance, we might produce insights on how to exploit the capabilities of AI while mitigating its weaknesses.

Review of lymph nodes is routinely performed after surgical removal of a tumour. In order to assess whether the tumour has spread to the lymphatic system, pathologists will check each lymph node for signs of metastasis. If the tumour has spread, the patient will be offered chemotherapy or other adjuvant therapies to increase life expectancy.

As part of the regular procedure following surgical removal of colorectal adenocarcinomas, at least 12 lymph nodes are extracted, cut into multiple thin sections, placed on glass slides and subjected to microscopic examination by a pathologist [27]. As part of this procedure to characterise lymphatic metastasis, the pathologist will need to identify the presence or absence of tumour, whether the tumour is the same as the primary resected tumour and note any additional findings. They will also count the total number of lymph nodes and report the positive and total counts.

Determining absence can be the most challenging part, sometimes necessitating systematic panning over the specimen in high resolution, as illustrated in fig. 6.4.

For tools to be considered safe, in general, they must not reduce the quality of decisions compared to manual methods. As for speed, a decrease might be motivated by a considerably higher level of quality, but even then, such a system will probably be hard to introduce into clinical practice since the domain is highly cost-averse. The main effect sought in the design was thus increasing the speed of decision making while maintaining or improving quality.

6.4.2 Process

Previous ML-related work on lymph node tumour detection has targeted sentinel lymph nodes from breast cancer patients [12], [113]. However, designing AI-assistance can be problematic for breast cancer patients due to the relative importance of detecting isolated tumour cells and several other low-prevalence findings besides tumour that any assisted approach risks omitting unless trained with an extensive and varied training data set. We instead chose to target lymph nodes harvested following resection of colorectal adenocarcinomas. The routine assessment does not prescribe identifying isolated tumour cells, and we believed that accidental findings, while still important, would be less of an issue.

An important early decision was whether to collaborate on the predictive model with another research group. In the end, we decided against it and chose to include the

Figure 6.4: Idealised panning path to conclude the absence of tumour. The tissue is viewed in 10X or 20X magnification as indicated by green boxes. The review follows the edge of the node capsule, then the interior is examined as shown by the yellow path.

data collection and model construction as part of the design exploration. The reasons were several, the major one wanting to study human involvement when constructing such systems “end to end”, and thus not presuppose any pre-defined boundaries between activities. Based on my previous engagements, we wanted to explore connections between the chosen interaction design, what the model is trained to predict and which data is gathered to train the model. Using a single, pre-trained model from another group would limit our possibilities to explore the holistic aspects of designing the joint system.

Similarly to the TissueWand exploration, a design-led approach was adopted 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.

The diagnostic process was characterised through field observations of pathologists performing the task as part of routine work. The observations included observing the workflow for the entire colorectal case from signout to reporting, not being limited to purely the lymph node assessment.

Chronologically, the exploration had two phases. The first was conducted between November 2017 and April 2018 and focused on collecting training data for an initial predictive model and then employing a rudimentary heatmap visualisation as a user interface that allowed evaluation of the model for the clinical task. The second phase took place between May 2019 and May 2020, focusing on designing a more sophisticated and valuable interactive system. Besides user-centred design work and interaction design,

the second phase also involved model development, additional data collection and a user study.

The initial phase involved one junior pathologist spending roughly 150 hours to annotate 436 slides. This relatively low-effort data collection effort yielded a model with a performance of 99.5%, as measured by the area under the receiving operating characteristic curve (AUROC). Operating at perfect sensitivity, that corresponds to an accuracy of 97.3%, which is similar to state-of-the-art works but still less than would be needed for autonomous decisions. How could the capability of the model be exploited despite its limitations?

Several guiding concepts for exploiting the model were considered. Our final concept combined two principles. First, when the model detects a positive region, the user's attention can be guided such that the search terminates quickly. Second, when there is no positive target, the search can terminate after reviewing a subset of the total specimen if 100% sensitivity is guaranteed.

When using a binary model, selecting the threshold for positivity involves a trade-off between *specificity* (number of false positives) and *sensitivity* (number of true positives found). Given an imperfect model, aiming to be 100% sensitive will produce false as well as true positives. The task for the user becomes rejecting false positives and accepting true ones. It was assumed that pathologists could perform this task with high accuracy if their attention was guided to a few selected areas. The benefit of such a strategy is that, when performed, there is no need to perform a subsequent manual systematic search.

During iterative prototyping, some inherent challenges with this strategy were identified. First, it must be fast in order to compete with manual review. Second, giving AI advice is known to cause automation bias, where users tend to over-rely on the prediction of the computer. Finally, users need to trust the strategy in order to actually use it, given that they have alternative methods to employ, such as a manual review. Additionally, guaranteeing high sensitivity is challenging because model behaviour can change due to changes in how images are produced, so-called data shifts [111].

In the prototyping process, we started with sketches of the concept that were later refined into static interface mockups. These initial sketches proved insufficient for eliciting how such an interface would work in use. While it was possible to convey success-path scenarios in which the pathologist was immediately guided to the tumour, it proved difficult to probe the less-than-ideal scenario and how the pathologist would act and experience the system on cases without positive areas. Hence we built a low-effort interactive prototype that faked many details of the prediction pipeline.

A fully functional system would require several components, including lymph node tissue segmentation, a tumour classifier, segmenting predictions into regions, and filtering which regions to present. In order to quickly assess the concept, the initial prototype faked all of these aspects. Using the heatmap from the previous phase, a few images containing normal and tumour-containing lymph nodes were selected. Then, prediction regions were manually produced based on heat map visualisations, creating a few scenarios where the model picked false positives or missed tumours. The lymph node tissue was manually labelled to simulate the segmentation step. Finally, the prototype was constructed with

Figure 6.5: Overview of the RAVS user interface. Lymph nodes are identified across the case and shown in a list of sections on the left. For each section a fixed number of regions are suggested for manual review.

functional navigation between lymph nodes and prediction areas but with pre-defined outputs instead of actual model predictions.

The prototype and the model were later refined over a few months using input from pathologists, adding detail to both interactive and computational aspects. For instance, we addressed how regions were visualised, how the region segmentation determined the proper size and magnification to show areas in, and improved the model’s capability to produce regions that were sensibly ranked in terms of the assigned confidence score. The initial use of a faked model allowed us to iteratively refine both the interaction and the model’s actual performance while still producing a coherent whole. This was helpful in order to gain feedback from pathologists during the process that could inform our actions.

When the iterative prototyping process had been going on for roughly eight months, we had gained the confidence to subject the system to a more extensive user study, comparing the AI-assisted work to manual work.

6.4.3 Resulting artefacts

The final prototype, Rapid Assisted Visual Search (RAVS), is a human-AI interactive tool used to efficiently perform lymph node reviews with maintained quality.

In an evaluation with six pathologists on 675 sections, RAVS improved the time on an average case from 17 minutes of manual review to 6 minutes. Quality was maintained, with a slightly higher sensitivity using RAVS. Additionally, we also showed that implicit collection of training data through use was capable of improving the prediction performance on a new set of unseen cases, increasing the AUC from 98.8% to 99.9%.

The user interface and the associated evaluation is described in more detail in Paper 2. The two datasets collected are available as AI-LNCO1 [82] and AI-LNCO2 [83].

Figure 6.6: Rejected prototype showing lymph node section without (left) and with a predictive heatmap (right). Movements such as panning or zooming in any of the two panes were synchronised. The user could navigate either pane, and movements would automatically be reflected in the opposite pane. Note that this case did not contain tumour, and intensive green areas on the right were false positives.

6.4.4 Main insights

Develop an explicit, synergistic strategy and then design for the transformed task. The design process used can be simplified as:

- Identify strengths and weaknesses of human and automatic approaches
- Find a suitable strategy that combines both model and human strengths
- Identify challenges of the transformed task
- Design a user interface that makes using the strategy efficient
- Describe metrics that a model that suits the strategy should have
- Optimise model training towards the needed characteristics

In our case, the suitable strategy was aiming for a 100% sensitive model and suggesting a fixed number of regions for pathologist review. However, the strategy was only iteratively discovered. At the end of the first phase, the ML system consisted of a dual-pane view with the original image to the left and a prediction heatmap to the right, as in fig. 6.6. The views were synchronised such that movement in one pane was reflected in the other. The dual-pane design was motivated by not wanting the heatmap to occlude the relevant features in the image for the human and to avoid a modal “toggle-style” design.

The evaluation of this early prototype was primarily qualitative, focused on discovering which qualities of the interaction the pathologist found important. Efficiency measurements pointed towards the pathologist being slower and less accurate when having the aid of the heatmap than using a manual process with no support at all. The pathologist reported that this version was unsuitable for clinical use due to the high level of false positives. While this might have seemed like a severe setback, the model was relatively capable of classifying tumours. We hypothesised that the effect might be due to an overemphasis on double-checking each potential detection of the model. The initial attempt had been quite naive, just presenting the model output in a data-intensive format. This led us to seek an explicit, safe strategy that the user could be guided to perform.

Figure 6.7: ML output as detailed regions (left) or fields-to-review (right). Viewing outputs as detailed regions became a distraction when borders appeared unintuitive.

Similarly, initial model development strived to produce models with a high AUC. Later, we primarily optimised towards achieving a high sensitivity given that a simulated perfect pathologist would consider the top-k regions.

In summary, initially, user interface development was guided by model capability, but in later stages, the way predictions were presented in the user interface changed the goals for model development.

Visualise ML output with respect to the information needed for the task. During the design process, the intent was re-framed from visualising potential detections to guiding the user to relevant fields of view. This re-framing impacted both computational and interactive considerations.

To illustrate this point, the output from an earlier prototype is presented next to the final version in fig. 6.7. Presenting outputs as detailed regions had several shortcomings. First, users became distracted from their task by the shape of the area. For an area that clearly contained tumour cells, when the region did not perfectly match their expectations, they lost time by pondering why some pixels were included while others were not. Second, when there were small but adjacent areas, they became annoyed at having to re-review an area that had already been part of an earlier field of view, even if only present in the periphery.

In order to accommodate viewing fields of view instead, the segmentation of positive pixels into areas was changed to be elastic, staying small only if the prediction was certain, otherwise growing, but with growth severely impacted if the area grew beyond what was visible in a single field of view. Finally, only the bounding box of the area was actually presented to the user in the form of a rectangular region.

In a similar vein, the confidence produced by the model for the region was not presented to the user in any way. The intent was that users should be unbiased when making the

decision for the region. However, as revealed through the user study presented in Paper 2, it might have been prudent to provide means for the user to discover decisions in conflict with the predictions. The preferred way in order to minimise bias would be to re-assess areas where their decisions contradicted the model at the end of the case.

Afford building trust progressively: expose residual uncertainty. Based on the model’s performance on a test set, in order to be 100% sensitive, we opted to show four regions to the user for each lymph node. However, would only reviewing four regions be safe? In general, a model’s performance will decline when deployed into actual clinical use. In addition, users might not trust the strategy until demonstrated and seek to gain personal experience that the strategy is safe.

In order to cover different scenarios of trust and predictive performance, we sought to make the sensitivity configurable but maintaining a minimal impact on reading time. The goal was that a case with many positive findings could be done quickly, but an entirely negative case might require reviewing more areas to gain confidence. Similarly, users sceptical to the system could use it with high sensitivity until they had gained enough trust to use it with a lower setting.

One way to control sensitivity is by letting users control the confidence threshold (0...1). The design process revealed that controlling this parameter did not correspond to anything useful for the pathologist. I.e., how safe would a value of 0.89 be? 0.78? In addition, when the threshold is controlled directly, the number of produced regions varies.

Instead, we adopted to let the pathologist control the number of regions shown in increments of four, with the default of starting at the lowest setting. The setting was controlled through a slider, as shown in fig. 6.8. At the highest setting of 20, given how the area segmentation algorithm determines fields of view, the entire lymph node is covered with regions. This way of controlling sensitivity provides a simple mental model for the user to reason about their current uncertainty. In the interviews following the user study, participants often referred to the number of regions they had viewed and the remaining unseen regions. e.g., “I have now reviewed four regions. I might go to eight just to be confident”, and “I reviewed many cases initially with eight regions, but true tumour was always in the first four”.

This interaction can be viewed as allowing the user to explore the model’s competence envelope, with the components of the user interface (i.e. regions) acting as a shared vocabulary along a shared, understandable dimension (up to the total area of the lymph node). The user support is not for a static level of appropriate trust but rather a dynamic, trusting relationship through repeated interactions over time. This is similar to motivation for the very rapid responses in TissueWand (Paper 1).

Adding a fixed threshold was considered, where some areas would be omitted, but there was a strong preference among pathologists for always having regions to go through, even for high-confidence normal lymph nodes that could theoretically be skipped.

Learning from use. Data collection is costly in terms of time and resources when performed up-front. However, the workflow changes introduced through an AI assistant can make additional data collection part of task completion. This simple and powerful

Figure 6.8: Sensitivity slider. Dragging the slider one step results in four more suggested regions. The intended use was to increase sensitivity if needed after going through the case once. Letting the slider control the number of suggested regions rather than a classification threshold makes the added effort predictable for the user.

idea has roots in the seminal work on human computation by von Ahn [124]. This work situates such considerations in high-stakes medical decisions by example. In essence, the design effort should aim at providing the proper incentives while maintaining minimal task impact in terms of time and effect on quality.

In theory, model updates could be made continuously. However, due to the safety-critical nature of the work, we imagine that model updates would be performed offline in order to be associated with additional quality assurance processes and enable communication to users about model changes.

Beyond producing better predictions, there is another potential synergy between task-driven, implicit data collection and iterative user interface prototyping. In model-focused research, data is considered fixed once collected. Typically, the data is separated into training, validation and holdout sets. Model training seeks to optimise a performance metric on the validation set through iterative experimentation on how to learn from the training data most effectively. Finally, the approach is evaluated on the holdout set once and only once.

When research involves both model and user interface iteration, if the user interface can collect training data in the prototype evaluation sessions, and unannotated data can be selected for each session, these sessions can have a similar role as the holdout set. The results of such a session are insights on the user experience, as well as new ground truths on which to evaluate the model. In addition, if the expectation is to repeat user studies, the data from previous iterations can be used as training data for the next session. This synergy between user studies and training datasets has also independently been suggested by Le et al. [66].

6.4.5 Concluding remarks

The exploration produced Rapid Assisted Visual Search, which is described together with a user evaluation in Paper 2. RAVS is one of few existing examples demonstrating how AI-assistants in digital pathology can safely reduce the workload of pathologists.

The associated paper did not describe the design process and intermittent insights, which motivated the longer process described in this section.

Our results indicate that designing human-AI interaction is not entirely different from traditional UI design. However, ML model limitations need to be proactively mitigated. We devised an explicit strategy for exploiting AI capabilities and designed the workflow to primarily support performing that strategy efficiently. Iterative prototyping was essential to discover details on how outputs should be filtered and visualised. Rather than designing for trust as a static state, we focused on the dynamic relationship for establishing trust, or trusting, through repeated interactions over time.

As mentioned in the paper, we also showed that AI-assistants such as RAVS could be designed to learn from use. This illustrates the far right of the dimension of designing for data capture put forth in Paper 3.

6.5 AI-assisted Quantification (PDL1)

6.5.1 Design situation

Pathologists routinely employ immunohistochemical stains to characterise or grade cancers. These chemical probes are engineered to enable visual assessment of specific gene expressions. The cells or part of cells where the targeted biomarker is present become coloured. Diagnostic protocols specify how this visual information is translated into reproducible characterisations. Typically, these protocols are correlated to treatment efficacy through clinical studies performed by drug manufacturers.

For this project, the intent was to design AI-support for a protocol that involves determining the ratio between positively and negatively stained tumour cell membranes. Specifically, we sought to support the protocol for determining whether PD-L1 inhibitor drugs are a viable treatment alternative for non-small cell lung cancer patients.

Some cancers hide from the immune system by a kind of cloaking mechanism and can effectively be treated by disabling the cancerous cells ability to do this. Recent studies have indicated breakthrough long-term survival results for eligible patients [34]. However, not all cancers hide by this mechanism and determining whether a patient shall receive this expensive treatment involves following comparatively new diagnostic procedures.

In short, cells are stained such that the cell membrane of cells having the cloaking ability becomes brown. According to the associated protocol, for treatment to be effective, the tissue needs more than 50% of the cancerous cells with a stained membrane. If the tissue has more than 1% stained cancerous cells, the treatment might be effective. If the ratio is below 1%, the treatment will likely not work.

In order to determine the ratio, the pathologist must estimate the overall pattern or individually count cells in a possibly large tissue area. Both strategies are associated with challenges. Estimating the overall pattern can be difficult if the case contains many heterogeneous areas or if the appearance is very close to one of the decision-cutoffs. On the other hand, individually counting cells is severely time-consuming and complicated by the presence of non-specific staining.

In contrast to my other design experiments, I was not present during the earliest phases of the design process. The characterisation of early design choices in the process is based on conversations with the lead designer, who was also present during the initial stage of my re-design.

As with the previously described project, the work was done in the context of a decision-making task, with the design goal of providing clinical utility through faster or higher quality outcomes.

6.5.2 Process

Understanding existing workflow. We used two sources to understand the existing workflow; protocol guides provided by the manufacturers of the associated immunohis-

tochemical staining kits (primarily Ventana SP263) and interviews with pathologists who routinely perform PD-L1 reviews.

The protocols recommend two strategies. First, they provide examples of how the pattern of various ratios appear visually and recommend using this approach when the ratio is clearly above 1%. For decisions near 1%, the protocol recommends counting individual cells. Pathologists similarly reported employing these two basic strategies. They can look at the overall impression of the image and use their experience to estimate the percentage right away. This is a swift decision but can be challenging if the tissue is close to the 1% or 50% cut-off levels. In addition, the strategy also becomes challenging if the pattern is heterogenous across the tissue.

Interestingly, pathologists reported reasoning about their confidence and working more slowly and carefully if they believed the decision was close to a cut-off. One pathologist reported having devised a strategy based on identifying major heterogeneous areas, determining the ratio in each area by visual pattern matching or heuristics and then calculating an average ratio weighted by the size of areas. Additionally, when still uncertain, the case would be independently scored by a second pathologist.

The second strategy involves manually counting both the tumour cells and the stained membranes and then deriving the ratio of the two. All things being equal, this second method will result in a relatively accurate decision, but it is orders of magnitude more time demanding. As a middle ground, pathologists sometimes choose a smaller area as a representative sample and only count within that area.

Initial strategy for AI-Assistance. A machine learning-based predictive model could always use the second strategy, making a decision at the cell level and reporting the exact ratio deriving from the two counts. We collected and manually annotated roughly 250 cases and then trained an ML model to perform this task.

If the model was perfect, a potential interaction could be that the user delineates an area and receive the result as a text summary, as depicted in fig. 6.9.

We identified several challenges with such an interaction, which are similar to the core issues for eXplainable AI: How does the expert know whether to trust the model? Our model had a cell-level accuracy of 85%, meaning it is occasionally wrong; how can the expert detect those instances? In essence, using the interface of fig. 6.9, the user has two options: either they accept the result blindly or reject it and perform their usual manual procedure.

An alternative might be visualising the result for each cell and letting the user verify and correct the cell if needed. However, since this might involve tens of thousands of cells, such an approach is clearly impractical.

In an attempt to make an intermediate level of automatic support, we designed a system where the user can explore the result by viewing and verifying a *systematic subset of decisions* of the underlying model, as depicted in fig. 6.10.

The user can increase the certainty of decisions by adding points, making the grid denser. This operation has been built such that adding points do not modify existing points since that would require the user to re-verify possibly verified points.

Figure 6.9: Illustration of model prediction visualisation with limited interactivity. The predictive model identifies all tumour cells and then counts the ratio between positively and negatively stained membranes. Only the final ratio is reported (shown in white text on green).

Feedback from pathologists indicated problems with this approach. While it produced good results for decisions far from decision cut-offs, the need for confidence on cases close to cut-offs necessitated such a high density of patches that the verification and correction effort again became impractical. In addition, the subset of cases where the prototype worked could easily be performed by the fast manual pattern matching strategy, making manual work faster than using this style of AI support.

Revised strategy; batching verification and correction We sought to remedy this by finding another intermediary level with more automatic support than the one above but less than only getting a final percentage. We analysed the bias-based error in our underlying model and found that most errors are somewhat systematic; visually similar patches might all be assigned the “wrong” classification, as depicted in fig. 6.11.

Based on this, we added an algorithm for unsupervised visual similarity clustering to our system and sought to design the interaction such that the user could work by only making decisions on a cluster level. This prototype is depicted in fig. 6.12.

In this revised prototype, the user can choose to look at the resulting percentage (top left), view the first patch of each cluster, or expand clusters and inspect their constituent patches. Additionally, clusters are ordered by uncertainty, and patches within the cluster are also ordered by certainty. The intent is that the user can detect errors by reviewing only the first few clusters and then accept the rest.

The workflow was as follows:

1. Review the overall assigned percentage. Is it reasonable given the overall look of the tissue?

Figure 6.10: Patch Gallery prototype. The predictions are patches sampled in a grid (top right) and can be interacted with either in a gallery of patches (left) or in the context of the tissue (right). The user is able to correct false classifications by clicking and dragging in the gallery, or by clicking a patch in the magnified main view (bottom, right).

Figure 6.11: t-SNE visualisation of clusters. We manually highlighted clusters that got the wrong prediction. In essence, they are close to each other (4 highlighted regions) and if we could cluster them, the user could batch-correct them.

Figure 6.12: Revised Patch Gallery. Patches are grouped by visually similarity in the left-most panel and can be batch-corrected. To the right, the predictions are visualised in the context of the tissue. Only patches on the grid are classified. Negative tumour cells appear with a green glyph, positive cells in orange. When zoomed in, the glyph changes such that the cell is not occluded (not shown).

2. Start reviewing the top-most clusters; based on its first patch, does it have the correct classification? If not, correct the classification. Observe the updated percentage and the confidence.
3. Check the patches in the cluster; does any patch “stand out” as not belonging to the cluster? Correct the patches by dragging them to the correct category; they will automatically be assigned to another cluster of that type.
4. Proceed through a few clusters; once no or few errors are detected, the rest is probably correct.

We hoped that pathologists would, through practice, learn to navigate this confidence-effort tradeoff. However, our user studies showed failures to complete the assigned tasks, reportedly due to a lack of motivation to go through the time-consuming learning process.

At the time of writing, we are still unable to say whether a time-efficient strategy could be learnt. However, it is clear that leaving so much control over the process and the confidence-effort tradeoff to the user is problematic. In interviews following the use of this prototype, pathologists expressed a desire to have more explicit guidance; exactly which patches should they review? How would their actions affect the certainty of the result? Overall, most were overwhelmed by the potentially huge number of patches to review. They struggled to understand how they should navigate the problem and iteratively make smaller decisions to gain the needed confidence. They thought the prototype could potentially be useful for tough cases but were uncertain of the usefulness for the majority of less challenging cases.

6.5.3 Main insights

While this exploration did not yield a user interface fulfilling the design goals, our experiences provided insights into designing for machine learning.

Designing for verification and correction: verification staircase. The recent rising interest in AI and neural networks have led to research into one of the major concerns with this type of machine learning: they result in incomprehensible “black box” models. Work that specifically addresses this issue has been called eXplainable AI (XAI), and research ranges from philosophical work on the nature of explanations to highly mathematical formulations of how the information in ML models is natured.

The design exploration of this project yielded a key idea that we believe can be useful in guiding other design efforts: that viewing phenomena at multiple, connected levels of detail might be considered a way of explaining. The argument and rationale for this design strategy are given in Paper 5.

Verification staircases might be a common set of principles to make ML-based decision support systems for situations where the model might not be capable of working autonomously.

We found that our “explanations” are all actionable; we think this is an important property of these designs. It is highly compatible with a human-centred perspective of viewing the partnership as an algorithm-in-the-loop and not human-in-the-loop. The effect stems not only from that the human receives an explanation, nor that she verifies results that

the machine has made, but that being able to act upon explanations creates a coupling of the two that achieves collaboration towards the same goal. Being able to drill down to the underlying phenomena at different fidelity acts as a shared common ground for communication.

6.5.4 Concluding remarks

The design experiment illustrates how designing useful human-AI interactions can be challenging due to the imperfect nature of predictive models and the difficulty of competing with specialists who, when unassisted, possess a repertoire of strategies that can flexibly be employed depending on the needed precision.

Nevertheless, the project provided insights into a connection between work on eXplainable AI and the role of designing for verification and correction in high-stakes medical decision making. Based on this, Verification Staircase (Paper 5) was suggested as a potential strategy for designing actionable explanations.

Chapter 7

A Framework for Designing Human-Centred Machine Learning

This chapter proposes a conceptual framework for guiding action when designing ML systems with a human-centred perspective. Specifically, the framework synthesises the activities and methods employed in the previous chapter into three cornerstone groups and highlights their interdependence.

Design involves utilising limited resources to discover constraints and shape solutions simultaneously. One key aspect of user-centred design is the importance of iterative design processes. By focusing on the relationship between the groups in the framework, the intent is to dynamically guide action that allow meaningful user-centred feedback throughout the design process without prescribing specific sequences or phases. Examples of where non-linear guidance might be needed include avoiding overinvestment in performance tuning, data collection or user interfaces concepts that might prove impossible to realise in practice.

7.1 The Importance of Thoughtful Action

Designing involves devising courses of action in order to purposefully move from a current situation towards a desirable future. However, what are the options to consider, and how does one choose one action over another? Core questions for design researchers working with machine learning are if and how human-centred design processes adapt to the arguably more emergent and unpredictable technological material that is machine learning.

The suggested framework is intended to guide thoughtful interaction design by focusing on the actions and methods that can be applied during different stages of an iterative design process. Using the framework as a scaffold, the insights provided in the previous chapter will be deepened to highlight how the framework can help identify challenges and decide between actions in order to resolve design choices.

The framework centres on three groups of related activities central to interaction design involving machine-learning components. The groups were identified by analysing the

Figure 7.1: Suggested framework for iterative ML system design. The framework relates three interconnected activities. *Data gathering* involves data collection, curation and labelling. *Model development* corresponds to algorithm and optimisation metric selection, training and hyperparameter tuning. Human-AI *interaction design* focuses primarily on designing the user interface, but also involves user-centred activities such as stakeholder interviews, workflow observations and user testing.

design processes of the design experiments mentioned in the previous chapter and filtered with regards to guiding answers to research questions on designing with machine learning identified in related work.

7.2 Three Interconnected Activities

When exploring the design space of an ML-based system, there is usually a trade-off between pursuing training algorithm improvements through *model development*, *gathering* more *training data* and attempting to *design* a suitable *interaction* using the current best performance of the predictive model. These activities are depicted schematically in fig. 7.1.

While the reality of practice is messy, this simple description of three highly ML-related activities can scaffold a discourse on how ML-specific constraints and possibilities affect key choices of the design process. Compared to the existing frameworks mentioned in sec. 3.2, it is intentionally simpler, containing only three overarching activities.

The design experiments described in Chapter 6 have typically begun in the corner of human-AI interaction design and involve user-centred design work. At some early point, one or more rather vague concepts are ideated with end-users. This process results in a problem definition, an idea of how ML might bring value and a very rough sketch

of how that value might be realised in use through interaction with end-users. Some quick and dirty model development follows to clarify what kind of training data to gather before a minimal data gathering pilot starts. After some added model development on that initial data, the ideal order of the design process becomes less clear. How much model development work is worth pursuing on a small amount of initial data? Should one blindly go collect more data? If so, how much? Furthermore, when should we revisit our human-AI interaction design to align concepts with new notions of achievable model performances?

Instead of prescribing some order of activities, it seems sensible that for each iteration, one can reflect upon the cost of each activity in relation to its predicted impact on advancing the overall design goal.

While designers probably have a good idea of the costs and benefits of employing user-centred design methods to further human-AI interaction designs given a fixed model performance, they might lack understanding of the effects and insights that could be achieved by data gathering and model development.

These effects will be exemplified in the following sections, focusing on the impact of training data on model performance, how user interface design can inform model development and how model development affects user interface design.

7.3 The Impact of Gathering Training Data

Recent work has highlighted that UX designers may not clearly understand the relationship ML has with data [35], thus elaborating how this relationship impacts design decisions might be generative in informing both processes and particular system designs.

Machine learning algorithms train a prediction model from samples. In general, the performance of the trained model improves with the amount and quality of training data. Since the model learns by incorporating new information, the value of each training sample will decline because, if drawn from the same source, chances increase that the new sample embodies something the model has already learned. This means that the generic shape of a learning curve follows an inverse power law, with the details depending on, e.g. model size, problem difficulty, and label noise [51]. It has been shown that this power law holds up to at least 300 million samples, as long as the accuracy level does not reach the inherent noise level in the data [117].

As described in Paper 3, a useful strategy is to first train an ML model on a relatively small dataset that has been expert-annotated by manual means, and then designing an interaction such that the collaborative workflow both aids the medical practitioner in her daily work while also implicitly generating new training data.

This dual-step strategy of ML system design can be viewed as designing to exploit the learning curve to address the cost of gathering data. Since the improvement to the model solely from data requires collecting data at increasing order of magnitudes, the costs of large-scale data collection prior to initial deployment might quickly grow out of proportion. Inversely, trying to deploy a product with too weak a model might not provide enough value to end-users to achieve the desired adoption.

The relationship between training data amount and model performance and its evolution over time, can thus be divided into an early *bootstrap* and a late *long tail* phase, with different characteristics and implications for design. At the bootstrap stage, the model improves rapidly with relatively small amounts of training data, and challenges include problem framing and training a model with sufficient performance to enable realistic user interface prototyping. At this stage, the design of manual annotation and labelling tools might be employed to engage domain experts as providers of training labels for the prototype models. The reasonable target performance in the bootstrap phase is probably less than that of the domain experts providing labels, if such are employed. Hence, the *human-AI interaction design* will need to consider that the error characteristics of the resulting model predictions can be compensated by human behaviour in the context of use.

When starting to approach the long tail, for many practical applications, further improvement to the model performance requires either the data collection to move towards large scale collection or attempting to improve the model by pursuing breakthrough research within machine learning. The design activities will focus on ways to make systems continuously collect training data so that it has minimal impact on end-users goals and user experience. In order to increase data collection by orders of magnitude, the collection will move from being explicit to implicit. Other notable issues the system design needs to address are model drift, quality assurance, systematic bias and generalizability between contexts.

Previous work has indicated that designers believe it may require an “unwieldy amount” of data to create functional prototypes [35]. The theory surrounding deep learning suggests it is possible to estimate the scaling of model performance with data if enough is gathered to get past a “small data” plateau and into the power-law region [51]. By combining experiments using small data and using the learning curve as a heuristic, designers can start imagining the points on the curve where resources required for data collection and model development exceed those that would be required to realise a system enabling valuable interactions with the current model performance. The designer could diverge to a few hypothetical designs by aiming for different target performance levels, as illustrated in fig. 7.2. If the prediction goals of those concepts are kept somewhat similar, this can be done with a low impact in terms of extra resources for data gathering and model development.

Creating a timeline for ML-based systems associated with machine performance and the amount of training data makes it possible to nuance questions such as *are human labellers useful or should we do something different?*. By the principles introduced, we might conclude that investing in labelling activities are particularly useful in the bootstrap phase, where time-consuming manual annotation work could be more fruitful than possibly overinvesting in more specific annotation tools. However, in later phases, the utility of this form of training data gathering declines, necessitating either highly efficient annotation tools or collecting training data implicitly.

Figure 7.2: Theoretical relation between model performance and training data amount. Knowledge of the relation can be used to proactively generate multiple design concepts adapted to near-term or long-term ML model performance.

7.4 The Impact of Interaction Design

The learning curve in fig. 7.2 relates the effect of additional training data to model performance with a fixed model training pipeline. However, model development activities such as algorithm selection, hyperparameter tuning, and devising training goal metrics also affect model performance. In the design experiments in this thesis, the explorations of this vast space of options were limited to applying techniques and pipelines that have been shown to be fruitful in similar scenarios or widely usable in general. The rationale behind this was based on two assumptions. First, similar to how the value of increased data declines, so does the impact of model development efforts. Second, the focus on human-AI interaction and the joint system means that the value of a slight increase in a performance metric is less important to designers.

With that stated, it does not mean model development is unimportant, but that it must be put into context in order to fruitfully decide what a reasonable amount of effort is and what should guide the process. For instance, the way regions are presented in RAVS affected both region segmentation and the learning goal. Specifically, since the order of predictions became important in the user interface, an entropy penalty was added to the loss function to ensure a broader range of confidence values.

The lack of user interface considerations can create issues at the very early stages of a process. For instance, in the DROID project mentioned in sec. 6.2, the lack of model development tasks for some efforts made data gathering considerations extremely hard to define.

7.5 The Impact of Model Development

Similarly to how a model development effort without explicit user interface intents become challenging to navigate due to unclear satisficing conditions, so does designing

user interfaces without knowing the limits and capabilities of ML models.

The design experiments in this thesis started with human-AI interaction design activities, such as observing and learning about current workflows and was followed by early concept sketches. Conceptual work without being informed by technological capabilities risks overspending on concepts that, while interesting, might be impossible or too resource-intensive to realise in practice. For RAVS (Paper 2), early concepts were loosely informed by the state of the art predictive models on similar tasks. One alternative, technology-rooted, ideation method suggested by others for bridging the gap between capabilities and potential user interfaces is a modified WoZ where the system's output is informed by a competent data scientist [131].

Once there is a preliminary ML model, activities used within model development to inform data gathering or algorithmic tuning can also inform user interface design. For instance, model debugging activities aim to explicate the errors that produced models makes. In the RAVS system, the central concept revolves around mitigating the lack of specificity when using a perfectly sensitive model. The number of fields of view to view, and whether this was a safe strategy was informed by model debugging activities in which perfect users were simulated. In the PDL1 project mentioned in sec. 6.5, t-SNE visualisations, a common form of model debugging, provided the insights that allowed us to aim for batch corrections through similarity clustering.

Chapter 8

Conclusion and Discussion

In this chapter, I summarise my contributions and discuss their implications on how we might design human-ML collaborative systems for efficient clinical decision making.

8.1 Summary of Contributions

The primary approach of this thesis has been constructive design research. The question of how we might design with machine learning has been explored through the making of prototypes, ML models and tools. The contributions made can be viewed from two perspectives, as contributions within pathology informatics and to the recent but rapidly growing field of human-AI interaction design.

Contributions to *pathology informatics*, primarily to image analysis in digital pathology, includes the *TissueWand* tool for rapid annotation (Paper 1) and *Rapid Assisted Visual Search (RAVS)* for lymph node metastasis detection (Paper 2). In particular, the positive impact of AI assistance demonstrated by RAVS is especially novel in that efficiency gains are made not only on target-present cases but also on target-absent cases. This is made possible by thoughtful interaction design and appropriate visualisation that allows a progressive build-up of trust. While the prototype for AI-assisted quantification of tumour ratios described in sec. 6.5 has not measurably demonstrated efficiency or quality improvements, the concept of a *verification staircase* (Paper 5) might be generative for future research. Visualising cell-level predictions on potentially tens of thousands of cells is impractical within a reasonable time frame. Thus, using a step-wise sampling strategy appears a viable alternative but need further research. Finally, the seven *datasets* produced can also be viewed as contributions to image analysis in digital pathology.

Designing efficient *human-AI interactions* involves more than merely focusing on the user interface. The design space of possibilities for supporting decision-making is partly defined by the qualities of the predictive models used, which depends on training data. If *designing* is defined as devising courses of action that can change existing situations into preferred ones, then the knowledge that can inform action potentially encompasses all human knowledge. Design space characterisations contribute by limiting the scope of potentially relevant knowledge. In Paper 3, I characterise how *the need for training data*

expands the role for design in both early and late phases of building ML systems supporting clinical decision making. In the position piece on the long tail of machine learning (Paper 6), I argue that knowing a few *principles on how predictive performance is affected by training data is vital for designers* in order to avoid pitfalls such as overinvesting in data collection or having the erroneous belief that problematic predictions on edge cases might disappear with more data alone.

Designers commonly utilise examples when generating ideas or concepts. In academic discourse, the emphasis is commonly on producing widely generalisable knowledge. This can create a disconnect between theory and exemplars. The lack of theory-related exemplars is especially noticeable in the growing field of human-AI interaction. One way of connecting theory with practice, and exemplars, is the notion of annotated portfolios [44], [78] produced by designers annotating their work. However, a shortcoming of annotated portfolios is precisely that they are limited to the designers' own work and thus overlook other potentially relevant sources of knowledge. As an alternative, in Paper 4, we suggested *curated collections* as a way to connect theory to exemplars by collecting and annotating exemplars around a set of salient traits.

Finally, TissueWand, Rapid Assisted Visual Search and our PD-L1 prototype all exemplify different ways in which the *capability uncertainty* and *output complexity* of machine learning is designed for at the decision-making user interface. Common to all three is that they approach verification and correction in a form tailored for the intended task. Likely, the specific interaction design used will only be re-usable for tasks with a similar situation structure [25]. Small changes in intent and context will require adaptation, which means that prescribing how to design for efficient clinical decision making needs to involve a method to guide the design process. A framework to guide designing for new human-ML collaborative systems was presented in Chapter 7. In contrast to existing process models for designing with ML, the framework does not prescribe a specific sequence of actions. Instead, the framework addresses how the more unfamiliar parts of designing with machine learning, namely data collection and model development, can relate to user interface design and be exploited at different phases of the design effort, depending on available resources and constraints of the problem space.

In the remaining sections of this chapter, I focus on positioning these contributions within the existing theory.

8.2 Why is Designing with ML Difficult?

Is human-AI interaction particularly challenging to design, and if so, why? One argument is that as a material, the properties of ML only emerge after training and that ML is thus a particularly shapeless material [115]. However, one might point out that many designerly materials have this property. Software is the material of interaction design, which has also been characterised in this fashion [80].

In another recent work, drawing on case studies, practitioner surveys, and teaching experiences, Yang et al. [133] suggested two attributes, *capability uncertainty* and *output complexity*, that contribute to difficulties reconciling user-centred design with machine

learning and AI. *Capability* refers to the functionalities of the resulting system (e.g., classifying membrane staining, detecting spam emails) and how well it can perform these functions, including the errors produced. In the early stages of design, the *uncertainty* of what is technically feasible makes ideas difficult to evaluate. Once a functional ML model has been built, analysing the performance and error characteristics of the model only provides partial information for design decisions due to biases in training data. When considering deployed ML systems that learn from use and continuously update the ML model, there is even more uncertainty. Finally, when considering interactive use, the range of use contexts, input data, errors produced and how users perceive those errors makes user-encountered capability highly uncertain. According to Yang et al. *output complexity* creates challenges due to the outputs that ML models produce, especially for pipelines combining multiple models. Specifically, realistic outputs cannot easily be simulated or mocked in order for the designer to gain insights on user-encountered capability. For instance, both simulations and human Wizard of Oz approaches would fail to capture how voice assistants such as Apple's Siri is experienced in use. Taken together, Yang argues, these two attributes challenge conventional design processes because those processes rely on designers being able to identify preferable futures without fully enacting each option.

Of particular interest to the work in this thesis, Yang argues that, from the perspective of *capability uncertainty* and *output complexity*, some human-AI interactions are not uniquely challenging for existing design methods. Such non-challenging systems include clinical decision support systems because they typically employ ML models that are static once deployed [129]. The argument is that understanding capability is straightforward: "*by playing with the system, the designers can develop a felt understanding of what the classifier can and cannot do. Because the system will not learn from new data, this understanding will remain valid post-deployment*" [133, p.8]. Furthermore, designers can easily imagine use scenarios where the functionality provides value, and because the outputs are limited (e.g. tumour, non-tumour), "*designers can enumerate all the ways the interaction may unfold*" [133, p.8].

Whether or not ML systems are "uniquely challenging" for existing design methods seems to be a rather academic question. My work does indicate, however, that even designing clinical decision support systems requires making and tinkering in addition to conventional lo-fi prototyping. Does this reliance on technological competence and time-consuming making defy the definition of design methods? It seems less important to answer this question and more essential to elicit insights pertaining to how to go about designing with ML.

The notions of *capability uncertainty* and *output complexity* capture important attributes for the design of ML systems — that these are at odds with the designer's ability to identify preferred future states. However, the claim that systems that rely on static, unchanging ML models considerably lessen the impact of these characteristics on the design process [129] does not fit well with my design experiments. In order to disentangle this apparent contradiction, we might consider explicitly adding the importance of possible future states to Yang's conceptual framework. For instance, we could distinguish between *future capability uncertainty* and *current capability uncertainty*. In addition, we could also be

more specific about the different perspectives from which capability can be evaluated. Capability solely determined by observing ML model outputs can be named *prediction capability* to distinguish it from *user-encountered capability*, which refers to the capability of the ML model combined with a user interface.

Using these extended concepts, we can begin to unpack the contradiction between the claim from Yang's framework and my design experiments. First, consider the early stages of a design process, where neither an ML model nor data might exist. Here *future capability uncertainty* is clearly at play, making ideas difficult to evaluate in terms of technical feasibility. At a later point, when an initial ML model has been created, my work and the framework proposed in the previous chapter suggests that the model should not be considered static nor final. Put differently, design-led work can inform both data gathering and model development which can affect future capability. The effect such actions will have on capability is associated with uncertainty. However, this does not mean that one needs to act blindly. I have suggested that some ML knowledge, such as the learning curve, can inform design decisions. Producing other similar heuristics or simplified ML knowledge is a prime candidate for future design research.

My work also indicates that having an existing ML model does not eliminate *current capability uncertainty*. The ML system can only be evaluated on available data, which means that post-deployment capability cannot be known unless all potential inputs are considered.

Finally, the two notions of *capability uncertainty* and *output complexity* do not address how to exploit the ML model's predictions despite its errors and imperfections. This is especially important for high-stakes, time-critical medical decision making. In my design experiments, it was never the case that I could enumerate all possible prediction errors and then simply design a solution—the impact on the overall user experience needed to be reconsidered for each tentative interaction or visualisation introduced. Imperfections or errors that were previously unproblematic might become nuisances or detrimental to the overall goal under a new interaction strategy. This is a consequence of the gap between *prediction capability* and *user-encountered capability*, bridged by the user interface design. The framework presented in the previous chapter explicates some of these relations and their role in design iterations: analysing an ML model's error characteristics can inform both data gathering and user interface design, but it is merely one step towards the desired future state, not the basis for a direct and immediate solution.

Future work along these lines might involve suggesting new theories that can provide analysis with more detailed descriptions of such situations. Adequately abstracted and theory-annotated exemplars might also work towards forming new *idioms* for user interfaces that deal with imperfect *prediction capabilities* in high-stakes decision making.

8.3 Power to the People? Reflections on Interactive Machine Learning

In my characterisation of the *machine teaching* design space [74] and my design experiments, there is little mention of interactive machine learning (IML) [38]. In this section, I will briefly discuss why this is so.

The core idea of IML is to allow users to interactively and iteratively train and then use their own ML models [36]. This is exemplified by toolkits such as Wekinator [40] and Regroup [7]. One reason to pursue IML is that ML models tend to be *narrow*: they are trained for one specific task. This is especially relevant for personal data, such as categorising email or webpages [62]. Allowing users to build their own models could extend the number of potential applications and democratise ML—giving “power to the people” [6]. The IML approach also presents a natural opportunity for merging the technical ML domain and design-related fields such as human-computer interaction.

In the earlier phases of my research, I imagined data-collection tools such as TissueWand [75] becoming part of a future IML toolkit for digital pathology. However, to understand what capabilities such a toolkit should have, I first wanted to understand a more traditional (i.e. non-IML) process for designing with ML. The framework in the previous chapter, while a simplification, summarises that understanding. To reiterate a few points, the actions one can pursue to do model development are vast and quickly evolving. Feedback from model development is needed in order to inform data gathering. Iteratively designing the user interface allows flexible adaptation to the produced ML model’s performance and the intended situation of use. This goes beyond getting more training data and employing more advanced training algorithms.

IML toolkits will, out of necessity, need to pre-define the user interfaces for interacting with ML model outputs and doing model development. The flexibility of IML tools for data gathering is countered by the limitations in the other two corners of my proposed HCML framework. The implication is that the range of potential IML-derived applications will be severely restricted by the pre-defined options implicit in the toolkit, which is not in line with the original vision of “power to the people”.

It is somewhat ironic that proponents of human-centred ways of designing ML systems wish to replace rich collaborations between individuals (e.g. domain experts, data scientists, designers) with one-stop software toolkits. Instead, one could imagine a shift towards ecosystems of tools and practices being thoughtfully applied according to need and context. This shift is partly in line with the vision of machine teaching [108] as a replacement for the IML umbrella. However, compared to the framework presented in this thesis, the definition of machine teaching lacks explicit attention to the user interface for interacting with ML outputs. Instead, the ML model seems to be the end-goal [99], with the final interaction to be added after model training. As remarked in the previous chapter, this neglects the potential of interaction design work to inform data gathering and model development.

Finally, as human-AI interaction design matures, the variability in both model development and user interface designs might converge on a set of processes, tools and interactive

idioms, justifying a revisiting of IML toolkits.

8.4 The Role of Constructive Design Research

In this section, I reflect on the merits of my chosen approach for resolving the research question of this thesis and whether the knowledge generated is valid and useful for designing human-AI interaction.

As outlined in sec. 4, the approach chosen for this work has been constructive design research (CDR). For the purpose of discussing the merits of the chosen approach, a particularly relevant comparison would be between CDR and conventional HCI research. Compared to research traditions in HCI, CDR is relatively young and actively being defined [60], [61]. One of many differences is that the epistemological roots of CDR are more influenced by pragmatism, and results are accepted to be provisional, contingent and aspirational [45] rather than analytically derived statements of truth. Reflective articulation and argumentative synthesis are more expected in CDR than in other, more objectivistic research traditions.

Due to its more objectivist roots, HCI research tends to emphasise empirical descriptions of how generalisable features of systems behave in experimental settings, along with cataloguing and describing usability issues. It is not uncommon to suggest guidelines, “dos and do not”, as an outcome of HCI research, one example being the recent guidelines of Amershi et al. [8]. However, descriptions of isolated features and guidelines give little guidance on how to construct ML systems tailored to novel contexts, nor do they sensitise designers to ML opportunities or extend the range of what is imaginable [94], [128].

The formulation of my research program started to take shape in 2015. At the time, few works were treating ML as a material for designers. Some influential exceptions were PhD theses from von Ahn [124], Amershi [4], and the seminal paper on explanatory debugging from Kulesza et al. [63]. However, none of these works adopted more recent ML techniques that avoid manual feature engineering, such as deep learning through convolutional neural networks. Within the medical imaging domain, human-AI interactions typically repeated common idioms of visualising predictions from ML models as heatmaps or as coloured overlays surrounding detection areas. Models of ML system engineering typically consisted of a linear sequence of steps, mostly ML model-oriented, with little guidance on how to adapt user interfaces to imperfect predictive capabilities or finding the right thing to design in the first place.

This knowledge gap was the reason for choosing a CDR approach; few examples to empirically investigate, few articulate accounts of design processes to generalise and little theory to use for analysis. To choose a design-driven approach entails, through making, starting to experiment with the available materials and actively reflecting upon what is brought to attention in the process. This attentive reflection leads to a growing, personal, material understanding. Through the process, one will generate artefacts that imbue key ideas. Together with those key ideas, the artefacts can be analysed and articulated as *generative* exemplars. Whether such accounts are to be considered scientific knowledge depends on whether the reader accepts the research tradition and its’ epistemologically pragmatic underpinning.

The description of novel tools and associated empirical user studies results presented in Paper 1 and Paper 2 are arguably more similar to traditional HCI research than, for instance, Paper 3 and Paper 5. The latter papers articulate key ideas derived from my design experiments, arguing for their generative value rather than solely relying on descriptions and empirical evidence. This type of knowledge contribution is concerned with what designing with machine learning *ought to be* and is mainly intended for other design researchers to build upon, to inform future theories or to inspire novel research artefacts.

Although future generativity is impossible to assess, the combination of tools and concepts presented in this thesis expands what designing with machine learning in digital pathology is and what it could be. The ML system for finding metastases in lymph nodes presented in Paper 2 exemplify how task-specific concern for how ML predictions are presented and interacted with can build the trust necessary to provide efficient assistance even in the absence of cancer. The design experiment on AI-assisted visual quantification described in sec. 6.5 guides future designers that seek to go beyond the accept-or-reject style of visualisations for quantification tasks.

All of the design experiments in this thesis have led to prototypes where the pathologist is clearly in control, verifying or overruling ML predictions. The interactions have been designed for being active and engaged, avoiding the boredom that can result from an overly supervisory role. In the ML systems investigated, there was no apparent need for the system to motivate its predictions further. The pathologist was assumed to be either more competent than any explanation that could be generated, or have other, better, means to resolve individual decisions. This competence dynamic between pathologists and ML systems might change as more training data and superior training algorithms become available. However, it is difficult to assess how such a change would affect the usefulness of the prototypes presented in this thesis. Even if ML systems become clearly superior to expert pathologists, the interactive in-the-loop style of interaction might still play an essential role in building trust and acting as a barrier against unanticipated prediction failures.

8.5 In Conclusion

This work has contributed to research on how we might design human-ML collaborative systems at the intersection of pathology informatics, human-computer interaction and design. Two tools that support clinical decision making and one tool to support interactive annotation were presented, together with qualitative or quantitative evaluations. Several datasets have been produced and have been made available to other researchers. Contributions to design include a characterisation of the design space when designing for teaching and a framework to guide action such that human-ML collaborative systems can be built in a human-centred fashion.

Bibliography

- [1] B. Acs, M. Rantalainen, and J. Hartman, “Artificial intelligence as the next step towards precision pathology,” *Journal of Internal Medicine*, vol. 288, no. 1, pp. 62–81, 2020.
- [2] C. Ahlberg and B. Shneiderman, “Visual information seeking: Tight coupling of dynamic query filters with starfield displays,” in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, (Boston, USA), ser. CHI ’94, New York, NY, USA: ACM, 1994, pp. 313–317.
- [3] O. Alvarado and A. Waern, “Towards Algorithmic Experience: Initial Efforts for Social Media Contexts,” in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, (Montreal, Canada), ser. CHI ’18, New York, NY, USA: ACM, 2018, pp. 1–12.
- [4] S. Amershi, “Designing for Effective End-User Interaction with Machine Learning,” PhD thesis, University of Washington, 2012.
- [5] S. Amershi, A. Begel, C. Bird, R. DeLine, H. Gall, E. Kamar, N. Nagappan, B. Nushi, and T. Zimmermann, “Software Engineering for Machine Learning: A Case Study,” in *Proceedings of the 41st International Conference on Software Engineering: Software Engineering in Practice*, (Montreal, Canada), Piscataway, NJ, USA: IEEE Press, 2019, pp. 291–300.
- [6] S. Amershi, M. Cakmak, W. B. Knox, and T. Kulesza, “Power to the People: The Role of Humans in Interactive Machine Learning,” *AI Magazine*, vol. 35, no. 4, pp. 105–120, 2014.
- [7] S. Amershi, J. Fogarty, and D. Weld, “Regroup: Interactive Machine Learning for On-demand Group Creation in Social Networks,” in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, (Austin, TX, USA), ser. CHI ’12, New York, NY, USA: ACM, 2012, pp. 21–30.
- [8] S. Amershi, D. Weld, M. Vorvoreanu, A. Fournery, B. Nushi, P. Collisson, J. Suh, S. Iqbal, P. N. Bennett, K. Inkpen, J. Teevan, R. Kikin-Gil, and E. Horvitz, “Guidelines for Human-AI Interaction,” in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, (Glasgow, Scotland, UK), ser. CHI ’19, New York, NY, USA: ACM, 2019, pp. 1–13.
- [9] A. I. D. Arena. (2021). AIDA Data Hub, [Online]. Available: <https://datahub.aida.scilifelab.se/> (visited on 07/12/2021).

- [10] M. Arvola and S. Holmlid, "Service Design Ways to Value-In-Use," in *Service Design Geographies. Proceedings of the ServDes.2016 Conference*, Linköping, Sweden: Linköping University Electronic Press, 2016, pp. 530–536.
- [11] L. Bainbridge, "Ironies of automation," *Automatica*, vol. 19, no. 6, pp. 775–779, 1983.
- [12] P. Bándi, O. Geessink, Q. Manson, M. Van Dijk, M. Balkenhol, M. Hermsen, B. Ehteshami Bejnordi, B. Lee, K. Paeng, A. Zhong, Q. Li, F. G. Zanjani, S. Zinger, K. Fukuta, D. Komura, V. Ovtcharov, S. Cheng, S. Zeng, J. Thagaard, A. B. Dahl, H. Lin, H. Chen, L. Jacobsson, M. Hedlund, M. Çetin, E. Halıcı, H. Jackson, R. Chen, F. Both, J. Franke, H. Küsters-Vandeveld, W. Vreuls, P. Bult, B. van Ginneken, J. van der Laak, and G. Litjens, "From Detection of Individual Metastases to Classification of Lymph Node Status at the Patient Level: The CAMELYON17 Challenge," *IEEE Transactions on Medical Imaging*, vol. 38, no. 2, pp. 550–560, 2019.
- [13] E. Beede, E. Baylor, F. Hersch, A. Iurchenko, L. Wilcox, P. Ruamviboonsuk, and L. M. Vardoulakis, "A Human-Centered Evaluation of a Deep Learning System Deployed in Clinics for the Detection of Diabetic Retinopathy," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, (Honolulu, HI, USA), ser. CHI '20, New York, NY, USA: ACM, 2020, pp. 1–12.
- [14] S. Bly and E. F. Churchill, "Design through matchmaking: Technology in search of users," *Interactions*, vol. 6, no. 2, pp. 23–31, 1999.
- [15] A. Bodén, J. F. Rose, M. Lindvall, and C. B. Stadler, *Breast data from the Visual Sweden project DROID*, 2019. [Online]. Available: <https://datahub.aida.scilifelab.se/10.23698/aida/drbr> (visited on 07/12/2021).
- [16] A. Bodén, J. F. Rose, M. Lindvall, and C. B. Stadler, *Colon data from the Visual Sweden project DROID*, 2019. [Online]. Available: <https://datahub.aida.scilifelab.se/10.23698/aida/drco> (visited on 07/12/2021).
- [17] A. C. S. Bodén, J. Molin, S. Garvin, R. A. West, C. Lundström, and D. Treanor, "The human-in-the-loop: An evaluation of pathologists' interaction with artificial intelligence in clinical practice," *Histopathology*, vol. n/a, no. n/a, 2021.
- [18] J. T. Browne, "Wizard of Oz Prototyping for Machine Learning Experiences," in *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, (Glasgow, Scotland, UK), ser. CHI EA '19, New York, NY, USA: ACM, 2019, pp. 1–6.
- [19] M. M. Bui, M. W. Riben, K. H. Allison, E. Chlipala, C. Colasacco, A. G. Kahn, C. Lacchetti, A. Madabhushi, L. Pantanowitz, M. E. Salama, R. L. Stewart, N. E. Thomas, J. E. Tomaszewski, and M. E. Hammond, "Quantitative Image Analysis of Human Epidermal Growth Factor Receptor 2 Immunohistochemistry for Breast Cancer," *Archives of pathology & laboratory medicine*, vol. 143, no. 10, pp. 1180–1195, 2019. pmid: 30645156.

- [20] W. Bulten, M. Balkenhol, J.-J. A. Belinga, A. Brilhante, A. Çakır, L. Egevad, M. Eklund, X. Farré, K. Geronatsiou, V. Molinié, G. Pereira, P. Roy, G. Saile, P. Salles, E. Schaafsma, J. Tschui, A.-M. Vos, H. van Boven, R. Vink, J. van der Laak, C. H.-v. der Kaa, and G. Litjens, “Artificial intelligence assistance significantly improves Gleason grading of prostate biopsies by pathologists,” *Modern Pathology*, vol. 34, pp. 660–671, 2021.
- [21] W. Bulten, H. Pinckaers, H. van Boven, R. Vink, T. de Bel, B. van Ginneken, J. van der Laak, C. Hulsbergen-van de Kaa, and G. Litjens, “Automated deep-learning system for Gleason grading of prostate cancer using biopsies: A diagnostic study,” *The Lancet Oncology*, vol. 21, no. 2, pp. 233–241, 2020.
- [22] C. J. Cai, E. Reif, N. Hegde, J. Hipp, B. Kim, D. Smilkov, M. Wattenberg, F. Viegas, G. S. Corrado, M. C. Stumpe, and M. Terry, “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), ser. CHI ’19, New York, NY, USA: ACM, 2019, pp. 1–14.
- [23] G. Campanella, M. G. Hanna, L. Geneslaw, A. Mirafior, V. W. K. Silva, K. J. Busam, E. Brogi, V. E. Reuter, D. S. Klimstra, and T. J. Fuchs, “Clinical-grade computational pathology using weakly supervised deep learning on whole slide images,” *Nature Medicine*, vol. 25, no. 8, pp. 1301–1309, 2019.
- [24] I. Cervin, J. Molin, and C. Lundström, “Improving the creation and reporting of structured findings during digital pathology review,” *Journal of Pathology Informatics*, vol. 7, no. 1, pp. 32–32, 2016.
- [25] E. K. Chiou and J. D. Lee, “Trusting Automation: Designing for Responsivity and Resilience,” *Human Factors*, p. 00 187 208 211 009 995, 2021. PMID: 33906505.
- [26] G. Cockton, “Designing Worth is Worth Designing,” in *Proceedings of the 4th Nordic Conference on Human-Computer Interaction: Changing Roles*, (Oslo, Norway), ser. NordiCHI ’06, New York, NY, USA: ACM, 2006, pp. 165–174.
- [27] 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, “Prognostic factors in colorectal cancer. College of American Pathologists Consensus Statement 1999,” *Archives of Pathology & Laboratory Medicine*, vol. 124, no. 7, pp. 979–994, 2000. PMID: 10888773.
- [28] A. Cooper, R. Reimann, and D. Cronin, *About Face 3: The Essentials of Interaction Design*. USA: John Wiley & Sons, Inc., 2007, 648 pp.
- [29] H. Cramer and J. Kim, “Confronting the Tensions Where UX Meets AI,” *Interactions*, vol. 26, no. 6, pp. 69–71, 2019.
- [30] N. Cross, “Designerly ways of knowing,” *Design Studies*, Special Issue Design Education, vol. 3, no. 4, pp. 221–227, 1982.

- [31] R. S. Crowley, G. J. Naus, J. Stewart, and C. P. Friedman, "Development of visual diagnostic expertise in pathology: An information-processing study," *Journal of the American Medical Informatics Association*, vol. 10, no. 1, pp. 39–51, 2003. PMID: 12509356.
- [32] S. W. A. Dekker and D. D. Woods, "MABA-MABA or Abracadabra? Progress on Human-Automation Co-ordination," *Cognition, Technology & Work*, vol. 4, no. 4, pp. 240–244, 2002.
- [33] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.
- [34] M. de Silva, M. Itchins, and N. Pavlakis, "Breakthrough 5-year survival with pembrolizumab in Keynote-001 study: Horizon shifting in advanced non-small cell lung cancer with immune check point inhibition," *Annals of Translational Medicine*, vol. 8, no. 8, p. 555, 2020. pmid: 32411778.
- [35] G. Dove, K. Halskov, J. Forlizzi, and J. Zimmerman, "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, Colorado, USA), ser. CHI '17, New York, NY, USA: ACM, 2017, pp. 278–288.
- [36] J. J. Dudley and P. O. Kristensson, "A Review of User Interface Design for Interactive Machine Learning," *ACM Trans. Interact. Intell. Syst.*, vol. 8, no. 2, 8:1–8:37, 2018.
- [37] M. Eiband, D. Buschek, and H. Hussmann, "How to Support Users in Understanding Intelligent Systems? Structuring the Discussion," in *26th International Conference on Intelligent User Interfaces*, (College Station, TX, USA), ser. IUI '21, New York, NY, USA: ACM, 2021, pp. 120–132.
- [38] J. A. Fails and D. R. Olsen, "Interactive machine learning," in *Proceedings of the 8th International Conference on Intelligent User Interfaces IUI 03*, (Miami Beach, FL, USA), ser. IUI '03, 2003, pp. 39–45.
- [39] M. Feinberg, "A Design Perspective on Data," in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, (Denver, Colorado, USA), ser. CHI '17, New York, NY, USA: ACM, 2017, pp. 2952–2963.
- [40] R. Fiebrink, P. R. Cook, and D. Trueman, "Human model evaluation in interactive supervised learning," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '11, New York, NY, USA: ACM, 2011, pp. 147–156.
- [41] J. L. Fine, "21 st century workflow: A proposal," *Journal of Pathology Informatics*, vol. 5, no. 1, p. 44, 2014. pmid: 25535592.
- [42] P. M. Fitts, "Human Engineering for an Effective Air-navigation and Traffic-control System," Ohio State University Research Foundation, Columbus, Ohio, Technical Report, 1951. [Online]. Available: <https://apps.dtic.mil/sti/citations/ADB815893>.

- [43] J. Forlizzi, I. Koskinen, P. Hekkert, and J. Zimmerman, "Let's get divorced: Pragmatic and critical constructive design research," in *Proceedings of IASDR 2017*, (Cincinnati, USA), 2017.
- [44] B. Gaver and J. Bowers, "Annotated portfolios," *Interactions*, vol. 19, no. 4, pp. 40–40, 2012.
- [45] W. Gaver, "What should we expect from research through design?" In *Proceedings of the 2012 ACM Annual Conference on Human Factors in Computing Systems - CHI '12*, ACM Press, 2012, pp. 937–937.
- [46] M. Gillies, R. Fiebrink, A. Tanaka, J. Garcia, F. Bevilacqua, A. Heloir, F. Nunnari, W. Mackay, S. Amershi, B. Lee, N. d'Alessandro, J. Tilmanne, T. Kulesza, and B. Caramiaux, "Human-Centred Machine Learning," in *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, (San Jose, California, USA), ser. CHI EA '16, New York, NY, USA: ACM, 2016, pp. 3558–3565.
- [47] F. Girardin and N. Lathia, "When User Experience Designers Partner with Data Scientists," Stanford University, Palo Alto, CA, Technical Report SS-17-04, 2017. [Online]. Available: <https://www.aaai.org/Library/Symposia/Spring/ss17-04.php>.
- [48] K. Goddard, A. Roudsari, and J. C. Wyatt, "Automation bias: A systematic review of frequency, effect mediators, and mitigators," *Journal of the American Medical Informatics Association*, vol. 19, no. 1, pp. 121–127, 2012.
- [49] B. Green and Y. Chen, "The Principles and Limits of Algorithm-in-the-Loop Decision Making," *Proceedings of the ACM on Human-Computer Interaction*, CSCW '19, vol. 3, 50:1–50:24, CSCW 2019.
- [50] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.
- [51] J. Hestness, S. Narang, N. Ardalani, G. Diamos, H. Jun, H. Kianinejad, M. M. A. Patwary, Y. Yang, and Y. Zhou. (2017). Deep Learning Scaling is Predictable, Empirically. arXiv: 1712.00409 [cs, stat], [Online]. Available: <http://arxiv.org/abs/1712.00409> (visited on 03/30/2019).
- [52] E. Hollnagel and D. D. Woods, "Cognitive Systems Engineering: New wine in new bottles," *International Journal of Man-Machine Studies*, vol. 18, no. 6, pp. 583–600, 1983.
- [53] L. E. Holmquist, "Intelligence on tap: Artificial intelligence as a new design material," *Interactions*, vol. 24, no. 4, pp. 28–33, 2017.
- [54] A. Holzinger, G. Langs, H. Denk, K. Zatloukal, and H. Müller, "Causability and explainability of artificial intelligence in medicine," *WIREs Data Mining and Knowledge Discovery*, vol. 9, no. 4, e1312, 2019.
- [55] M. K. Hong, A. Fourney, D. DeBellis, and S. Amershi, "Planning for Natural Language Failures with the AI Playbook," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, (Yokohama, Japan), ser. CHI '21, New York, NY, USA: ACM, 2021, pp. 1–11.

- [56] K. Höök and J. Löwgren, “Strong concepts,” *ACM Transactions on Computer-Human Interaction*, vol. 19, no. 3, pp. 1–18, 2012.
- [57] E. Horvitz, “Principles of mixed-initiative user interfaces,” in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, (Pittsburgh, Pennsylvania, USA), ser. CHI ’99, New York, NY, USA: ACM, 1999, pp. 159–166.
- [58] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, “Densely Connected Convolutional Networks,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2261–2269.
- [59] S. Khairat, D. Marc, W. Crosby, and A. Al Sanousi, “Reasons For Physicians Not Adopting Clinical Decision Support Systems: Critical Analysis,” *JMIR Medical Informatics*, vol. 6, no. 2, 2018. PMID: 29669706.
- [60] I. Koskinen, J. Zimmerman, T. Binder, J. Redstrom, and S. Wensveen, *Design Research Through Practice: From the Lab, Field, and Showroom*, 1st. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2012.
- [61] P. G. Krogh and I. Koskinen, *Drifting by Intention: Four Epistemic Traditions from within Constructive Design Research*, ser. Design Research Foundations. Cham: Springer International Publishing, 2020.
- [62] T. Kulesza, S. Amershi, R. Caruana, D. Fisher, D. Charles, T. Kulesza, S. Amershi, R. Caruana, D. Fisher, and D. Charles, “Structured labeling for facilitating concept evolution in machine learning,” in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, (Toronto, Ontario, Canada), ser. CHI ’14, ACM, 2014, pp. 3075–3084.
- [63] T. Kulesza, M. Burnett, W.-K. Wong, and S. Stumpf, “Principles of Explanatory Debugging to Personalize Interactive Machine Learning,” in *Proceedings of the 20th International Conference on Intelligent User Interfaces*, (Atlanta, Georgia, USA), ser. IUI ’15, New York, NY, USA: ACM, 2015, pp. 126–137.
- [64] A.-Y. Kwon, H. Y. Park, J. Hyeon, S. J. Nam, S. W. Kim, J. E. Lee, J.-H. Yu, S. K. Lee, S. Y. Cho, and E. Y. Cho, “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*, vol. 14, no. 2, pp. 1–13, 2019.
- [65] V. Lai and C. Tan, “On Human Predictions with Explanations and Predictions of Machine Learning Models: A Case Study on Deception Detection,” in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, ser. FAT* ’19, New York, NY, USA: ACM, 2019, pp. 29–38.
- [66] H. V. Le, S. Mayer, and N. Henze, “Deep learning for human-computer interaction,” *Interactions*, vol. 28, no. 1, pp. 78–82, 2020.
- [67] K. Lindman, J. F. Rose, M. Lindvall, C. Lundstrom, and D. Treanor, “Annotations, ontologies, and whole slide images – Development of an annotated ontology-driven whole slide image library of normal and abnormal human tissue,” *Journal of Pathology Informatics*, vol. 10, no. 1, p. 22, 2019.

- [68] K. Lindman, J. F. Rose, M. Lindvall, and C. B. Stadler, *Ovary data from the Visual Sweden project DROID*, 2019. [Online]. Available: <https://datahub.aida.scilifelab.se/10.23698/aida/drov> (visited on 07/12/2021).
- [69] K. Lindman, J. Rose, M. Lindvall, and C. B. Stadler, *Skin data from the Visual Sweden project DROID*, 2019. [Online]. Available: <https://datahub.aida.scilifelab.se/10.23698/aida/drsk> (visited on 07/12/2021).
- [70] M. Lindvall, C. Lundström, and J. Löwgren, “Rapid Assisted Visual Search: Supporting Digital Pathologists with Imperfect AI,” in *26th International Conference on Intelligent User Interfaces*, (College Station, TX, USA), ser. IUI ’21, New York, NY, USA: ACM, 2021, pp. 504–513.
- [71] M. Lindvall and J. Molin, “Designing for the Long Tail of Machine Learning,” in *Workshop on Emerging Perspectives in Human-Centered Machine Learning (CHI ’19)*, Glasgow, Scotland, UK, 2019. arXiv: 2001.07455.
- [72] M. Lindvall and J. Molin, “Verification Staircase : A Design Strategy for Actionable Explanations,” in *Proceedings of Workshop on Explainable Smart Systems for Algorithmic Transparency in Emerging Technologies*, vol. 2582, CEUR-WS.org, 2020.
- [73] M. Lindvall, J. Molin, and J. Löwgren, “From Machine Learning to Machine Teaching: The Importance of UX,” *Interactions*, vol. 25, no. 6, pp. 52–57, 2018.
- [74] M. Lindvall, J. Molin, and J. Löwgren, “The Importance of UX for Machine Teaching,” in *2018 AAAI Spring Symposium Series*, 2018.
- [75] M. Lindvall, A. Sanner, F. Petré, K. Lindman, D. Treanor, C. Lundström, and J. Löwgren, “TissueWand, a Rapid Histopathology Annotation Tool,” *Journal of Pathology Informatics*, vol. 11, 2020. pmid: 33042606.
- [76] G. Litjens, P. Bandi, B. Ehteshami Bejnordi, O. Geessink, M. Balkenhol, P. Bult, A. Halilovic, M. Hermsen, R. van de Loo, R. Vogels, Q. F. Manson, N. Stathonikos, A. Baidoshvili, P. van Diest, C. Wauters, M. van Dijk, and J. van der Laak, “1399 H&E-stained sentinel lymph node sections of breast cancer patients: The CAMELYON dataset,” *GigaScience*, vol. 7, no. 6, giy065, 2018.
- [77] Y. Liu, K. Gadepalli, M. Norouzi, G. E. Dahl, T. Kohlberger, A. Boyko, S. Venugopalan, A. Timofeev, P. Q. Nelson, G. S. Corrado, J. D. Hipp, L. Peng, and M. C. Stumpe. (2017). Detecting Cancer Metastases on Gigapixel Pathology Images. arXiv: 1703.02442 [cs], [Online]. Available: <http://arxiv.org/abs/1703.02442> (visited on 02/04/2021).
- [78] J. Löwgren, “Annotated portfolios and other forms of intermediate-level knowledge,” *Interactions*, vol. 20, no. 1, pp. 30–30, 2013.
- [79] J. Löwgren, “On the Significance of Making in Interaction Design Research,” *Interactions*, vol. 23, no. 3, pp. 26–33, 2016.
- [80] J. Löwgren and E. Stolterman, *Thoughtful Interaction Design: A Design Perspective on Information Technology*. The MIT Press, 2007.

- [81] Z. Lu and M. Yin, “Human Reliance on Machine Learning Models When Performance Feedback is Limited: Heuristics and Risks,” in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, (Yokohama, Japan), ser. CHI ’21, New York, NY, USA: ACM, 2021, pp. 1–16.
- [82] G. Maras, M. Lindvall, and C. Lundström, *Regional lymph node metastasis in colon adenocarcinoma*, 2019. [Online]. Available: <https://datasets.aida.scilifelab.se/10.23698/aida/lnc0> (visited on 02/14/2021).
- [83] G. Maras, M. Lindvall, and C. Lundström, *Regional lymph node metastasis in colon adenocarcinoma, second collection series*, 2020. [Online]. Available: <https://datasets.aida.scilifelab.se/10.23698/aida/lnc02> (visited on 02/14/2021).
- [84] S. McGrath, P. Mehta, A. Zyteck, I. Lage, and H. Lakkaraju. (2020). When Does Uncertainty Matter?: Understanding the Impact of Predictive Uncertainty in ML Assisted Decision Making. arXiv: 2011.06167 [cs], [Online]. Available: <http://arxiv.org/abs/2011.06167> (visited on 04/29/2021).
- [85] E. Mercan, L. G. Shapiro, T. T. Brunyé, D. L. Weaver, and J. G. Elmore, “Characterizing Diagnostic Search Patterns in Digital Breast Pathology: Scanners and Drillers,” *Journal of Digital Imaging*, vol. 31, no. 1, pp. 32–41, 2018.
- [86] T. Miller, P. Howe, and L. Sonenberg. (2017). Explainable AI: Beware of Inmates Running the Asylum Or: How I Learnt to Stop Worrying and Love the Social and Behavioural Sciences. arXiv: 1712.00547 [cs], [Online]. Available: <http://arxiv.org/abs/1712.00547> (visited on 12/19/2019).
- [87] M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, and T. Gebru, “Model Cards for Model Reporting,” in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, ser. FAT* ’19, New York, NY, USA: ACM, 2019, pp. 220–229.
- [88] J. Molin, “Diagnostic Review with Digital Pathology: Design of digital tools for routine diagnostic use,” PhD thesis, Chalmers University of Technology, Göteborg, Sweden, 2016. [Online]. Available: <http://publications.lib.chalmers.se/publication/244438-diagnostic-review-with-digital-pathology-design-of-digital-tools-for-routine-diagnostic-use>.
- [89] J. Molin, M. Fjeld, C. Mello-Thoms, and C. Lundström, “Slide navigation patterns among pathologists with long experience of digital review,” *Histopathology*, vol. 67, no. 2, pp. 185–92, 2015. PMID: 25487230.
- [90] J. Molin, P. W. Woźniak, C. Lundström, D. Treanor, and M. Fjeld, “Understanding Design for Automated Image Analysis in Digital Pathology,” in *Proceedings of the 9th Nordic Conference on Human-Computer Interaction - NordiCHI ’16*, ACM Press, 2016, pp. 1–10.

- [91] S. T. Mueller, R. R. Hoffman, W. Clancey, A. Emrey, and G. Klein. (2019). Explanation in Human-AI Systems: A Literature Meta-Review, Synopsis of Key Ideas and Publications, and Bibliography for Explainable AI. arXiv: 1902.01876 [cs.AI], [Online]. Available: <https://arxiv.org/abs/1902.01876v1> (visited on 12/18/2019).
- [92] C. Noessel, *Designing Agentive Technology: AI That Works for People*. Brooklyn, NY: Rosenfeld Media LLC, 2017.
- [93] D. A. Norman, "How might people interact with agents," *Communications of the ACM*, vol. 37, no. 7, pp. 68–71, 1994.
- [94] D. A. Norman, "Technology first, needs last: The research-product gulf," *Interactions*, vol. 17, no. 2, pp. 38–42, 2010.
- [95] R. Parasuraman, T. B. Sheridan, and C. D. Wickens, "A model for types and levels of human interaction with automation," *IEEE transactions on systems, man, and cybernetics. Part A, Systems and humans: a publication of the IEEE Systems, Man, and Cybernetics Society*, vol. 30, no. 3, pp. 286–297, 2000.
- [96] P. Pirolli and S. Card, "The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis," *Proceedings of International Conference on Intelligence Analysis*, vol. 2005, pp. 2–4, 2005.
- [97] B. Pohn, M. Kargl, R. Reihs, A. Holzinger, K. Zatloukal, and H. Müller, "Towards a Deeper Understanding of How a Pathologist Makes a Diagnosis: Visualization of the Diagnostic Process in Histopathology," in *2019 IEEE Symposium on Computers and Communications (ISCC)*, 2019, pp. 1081–1086.
- [98] R. Porter, J. Theiler, and D. Hush, "Interactive machine learning in data exploitation," *Computing in Science and Engineering*, vol. 15, no. 5, pp. 12–20, 2013.
- [99] G. Ramos, C. Meek, P. Simard, J. Suh, and S. Ghorashi, "Interactive machine teaching: A human-centered approach to building machine-learned models," *Human-Computer Interaction*, vol. 35, no. 5-6, pp. 413–451, 2020.
- [100] G. Ramos, J. Suh, S. Ghorashi, C. Meek, R. Banks, S. Amershi, R. Fiebrink, A. Smith-Renner, and G. Bansal, "Emerging Perspectives in Human-Centered Machine Learning," in *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, ser. CHI EA '19, New York, NY, USA: ACM, 2019, pp. 1–8.
- [101] R. Randell, R. A. Ruddle, P. Quirke, R. G. Thomas, and D. Treanor, "Working at the microscope: Analysis of the activities involved in diagnostic pathology," *Histopathology*, vol. 60, no. 3, pp. 504–510, 2012.
- [102] R. A. Ruddle, R. G. Thomas, R. Randell, P. Quirke, and D. Treanor, "The Design and Evaluation of Interfaces for Navigating Gigapixel Images in Digital Pathology," *ACM Transactions on Computer-Human Interaction*, vol. 23, no. 1, pp. 1–29, 2016.
- [103] M. Sendak, M. C. Elish, M. Gao, J. Futoma, W. Ratliff, M. Nichols, A. Bedoya, S. Balu, and C. O'Brien, "'The human body is a black box': Supporting clinical decision-making with deep learning," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, ser. FAT* '20, New York, NY, USA: ACM, 2020, pp. 99–109.

- [104] R. Sheline and C. J. MacLellan, “Investigating Machine-Learning Interaction with Wizard-of-Oz Experiments,” in *Learning by Instruction Workshop (NIPS 2018)*, ser. NIPS ’18, Montreal, Canada, 2018, p. 8.
- [105] T. B. Sheridan, “Comments on “Issues in Human–Automation Interaction Modeling: Presumptive Aspects of Frameworks of Types and Levels of Automation” by David B. Kaber,” *Journal of Cognitive Engineering and Decision Making*, vol. 12, no. 1, pp. 25–28, 2018.
- [106] T. B. Sheridan and W. L. Verplank, “Human and computer control of undersea teleoperators,” Massachusetts Institute of Technology, Cambridge, MA, Technical Report, 1978. [Online]. Available: <http://oai.dtic.mil/oai/oai?verb=getRecord&metadataPrefix=html&identifier=ADA057655>.
- [107] B. Shneiderman, “Human-Centered Artificial Intelligence: Reliable, Safe & Trustworthy,” *International Journal of Human–Computer Interaction*, vol. 36, no. 6, pp. 495–504, 2020.
- [108] P. Y. Simard, S. Amershi, D. M. Chickering, A. E. Pelton, S. Ghorashi, C. Meek, G. Ramos, J. Suh, J. Verwey, M. Wang, and J. Wernsing, “Machine Teaching: A New Paradigm for Building Machine Learning Systems,” Microsoft Research Lab, Redmond, WA, Technical Report MSR-TR-2017-26, 2017. arXiv: 1707.06742. [Online]. Available: <http://arxiv.org/abs/1707.06742>.
- [109] H. A. Simon, *The Sciences of the Artificial*, 3. ed. Cambridge, Mass.: MIT Press, 1968, 231 pp.
- [110] C. Sommer, C. Straehle, U. Köthe, and F. A. Hamprecht, “Ilastik: Interactive learning and segmentation toolkit,” in *2011 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, 2011, pp. 230–233.
- [111] K. Stacke, G. Eilertsen, J. Unger, and C. Lundström, “A closer look at domain shift for deep learning in histopathology,” in *MICCAI 2019 Computational Pathology Workshop COMPAY*, ser. MICCAI 2019 Workshop COMPAY, 2019.
- [112] C. B. Stadler, M. Lindvall, C. Lundström, A. Bodén, K. Lindman, J. Rose, D. Treanor, J. Blomma, K. Stacke, N. Pinchaud, M. Hedlund, F. Landgren, M. Woisetschlager, and D. Forsberg, “Proactive Construction of an Annotated Imaging Database for Artificial Intelligence Training,” *Journal of Digital Imaging*, vol. 34, no. 1, pp. 105–115, 2021.
- [113] D. F. Steiner, R. MacDonald, Y. Liu, P. Truszkowski, J. D. Hipp, C. Gammage, F. Thng, L. Peng, and M. C. Stumpe, “Impact of Deep Learning Assistance on the Histopathologic Review of Lymph Nodes for Metastatic Breast Cancer,” *The American Journal of Surgical Pathology*, vol. 42, no. 12, p. 1636, 2018.
- [114] H. Subramonyam, C. Seifert, and E. Adar, “ProtoAI: Model-Informed Prototyping for AI-Powered Interfaces,” in *26th International Conference on Intelligent User Interfaces*, ser. IUI ’21, New York, NY, USA: ACM, 2021, pp. 48–58.

- [115] H. Subramonyam, C. Seifert, and E. Adar, "Towards A Process Model for Co-Creating AI Experiences," in *Designing Interactive Systems Conference 2021*, New York, NY, USA: ACM, 2021, pp. 1529–1543.
- [116] S. Sugiyama, N. Yuhara, and S. Horiuchi, "The Effects of Participatory Mode on the detection of Dynamic System Failure," *IFAC Proceedings Volumes*, vol. 21, no. 5, pp. 279–283, 1988.
- [117] C. Sun, A. Shrivastava, S. Singh, and A. Gupta, "Revisiting Unreasonable Effectiveness of Data in Deep Learning Era," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 843–852.
- [118] S. Thorstenson, J. Molin, and C. Lundström, "Implementation of large-scale routine diagnostics using whole slide imaging in Sweden: Digital pathology experiences 2006-2013," *Journal of Pathology Informatics*, vol. 5, no. 1, p. 14, 2014. pmid: 24843825.
- [119] S. Tonekaboni, S. Joshi, M. D. McCradden, and A. Goldenberg, "What Clinicians Want: Contextualizing Explainable Machine Learning for Clinical End Use," in *Machine Learning for Healthcare Conference*, PMLR, 2019, pp. 359–380.
- [120] D. Tran Luciani, "Designing for sketching to support concept exploration," PhD thesis, Linköping University, Linköping, 2019. [Online]. Available: <http://public.eblib.com/choice/PublicFullRecord.aspx?p=5975125> (visited on 12/16/2019).
- [121] D. Tran Luciani, M. Lindvall, and J. Löwgren, "Machine learning as a design material: A curated collection of exemplars for visual interaction," in *DS 91: Proceedings of NordDesign 2018*, (Linköping, Sweden), The Design Society, 2018, pp. 1–10.
- [122] D. Tran Luciani, J. Löwgren, and J. Lundberg, "Designing fine-grained interactions for automation in air traffic control," *Cognition, Technology & Work*, 2019.
- [123] P. van Allen, "Prototyping ways of prototyping AI," *Interactions*, vol. 25, no. 6, pp. 46–51, 2018.
- [124] L. von Ahn, "Human Computation," PhD thesis, Carnegie Mellon University, 2005.
- [125] K. Wagstaff. (2012). Machine Learning that Matters. arXiv: 1206.4656 [cs, stat], [Online]. Available: <http://arxiv.org/abs/1206.4656> (visited on 02/01/2018).
- [126] X. Wang and M. Yin, "Are Explanations Helpful? A Comparative Study of the Effects of Explanations in AI-Assisted Decision-Making," in *26th International Conference on Intelligent User Interfaces*, ser. IUI '21, New York, NY, USA: ACM, 2021, pp. 318–328.
- [127] A. Wong, E. Otles, J. P. Donnelly, A. Krumm, J. McCullough, O. DeTroyer-Cooley, J. Pestrue, M. Phillips, J. Konye, C. Penozza, M. Ghous, and K. Singh, "External Validation of a Widely Implemented Proprietary Sepsis Prediction Model in Hospitalized Patients," *JAMA Internal Medicine*, 2021. pmid: 34152373.

- [128] Q. Yang, “Machine Learning as a UX Design Material: How Can We Imagine Beyond Automation, Recommenders, and Reminders?” Stanford, San Francisco, Technical Report SS-18, 2018, p. 6. [Online]. Available: <https://aaai.org/ocs/index.php/SSS/SSS18/paper/view/17471>.
- [129] Q. Yang, “Profiling Artificial Intelligence as a Material for User Experience Design,” PhD thesis, Carnegie Mellon University, Pittsburgh, Pennsylvania, 2020, 126 pp. [Online]. Available: https://kilthub.cmu.edu/articles/thesis/Profiling_Artificial_Intelligence_as_a_Material_for_User_Experience_Design/14376731.
- [130] Q. Yang, “The Role of Design in Creating Machine-Learning-Enhanced User Experience,” Stanford University, Palo Alto, CA, Technical Report SS-17-04, 2017, pp. 406–411. [Online]. Available: <http://aaai.org/ocs/index.php/SSS/SSS17/paper/view/15363/14575>.
- [131] Q. Yang, J. Cranshaw, S. Amershi, S. T. Iqbal, and J. Teevan, “Sketching NLP: A Case Study of Exploring the Right Things To Design with Language Intelligence,” in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, (Glasgow, Scotland, UK), ser. CHI ’19, New York, NY, USA: ACM, 2019, pp. 1–12.
- [132] Q. Yang, A. Scuito, J. Zimmerman, J. Forlizzi, and A. Steinfeld, “Investigating How Experienced UX Designers Effectively Work with Machine Learning,” in *Proceedings of the 2018 Designing Interactive Systems Conference*, ser. DIS ’18, New York, NY, USA: ACM, 2018, pp. 585–596.
- [133] Q. Yang, A. Steinfeld, C. Rosé, and J. 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*, ser. CHI ’20, New York, NY, USA: ACM, 2020, pp. 1–13.
- [134] Q. Yang, A. Steinfeld, and J. Zimmerman, “Unremarkable AI: Fitting Intelligent Decision Support into Critical, Clinical Decision-Making Processes,” in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, (Glasgow, Scotland Uk), ser. CHI ’19, New York, NY, USA: ACM, 2019, 238:1–238:11.
- [135] Y. Zhang, Q. V. Liao, and R. K. E. Bellamy, “Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making,” in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, ser. FAT* ’20, New York, NY, USA: ACM, 2020, pp. 295–305.
- [136] S. Zhou, H. Marklund, O. Blaha, M. Desai, B. Martin, D. Bingham, G. J. Berry, E. Gomulia, A. Y. Ng, and J. Shen, “Deep learning assistance for the histopathologic diagnosis of *Helicobacter pylori*,” *Intelligence-Based Medicine*, vol. 1-2, p. 100 004, 2020.
- [137] Z. Zhou, L. Sun, Y. Zhang, X. Liu, and Q. Gong, “ML Lifecycle Canvas: Designing Machine Learning-Empowered UX with Material Lifecycle Thinking,” *Human-Computer Interaction*, vol. 35, no. 5-6, pp. 362–386, 2020.

-
- [138] J. Zimmerman, J. Forlizzi, and S. Evenson, “Research Through Design As a Method for Interaction Design Research in HCI,” in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, (San Jose, California, USA), ser. CHI '07, New York, NY, USA: ACM, 2007, pp. 493–502.

Part II

Appended papers

Paper 1

TissueWand, a Rapid Histopathology Annotation Tool.

M. Lindvall, A. Sanner, F. Petré, K. Lindman, D. Treanor, C. Lundström, and J.Löwgren

Journal of Pathology Informatics 11(27), 2020.

Appended papers have been removed from the electronic thesis.

Available from: doi:10.4103/jpi.jpi_5_20

Paper 2

Rapid Assisted Visual Search: Supporting Digital Pathologists with Imperfect AI.

M. Lindvall, C. Lundström, and J. Löwgren

*In 26th International Conference on Intelligent User Interfaces
(IUI '21), 2021.*

Appended papers have been removed from the electronic thesis.

Available from: doi:10.1145/3397481.3450681

Paper 3

From Machine Learning to Machine Teaching: The Importance of UX.

M. Lindvall, J. Molin, and J. Löwgren

Interactions, 25(6), 52-57, 2018.

Appended papers have been removed from the electronic thesis.

Available from: [doi:10.1145/3282860](https://doi.org/10.1145/3282860)

Paper 4

Machine Learning as a Design Material: a Curated Collection of Exemplars for Visual Interaction.

D. Tran Luciani, M. Lindvall, and J. Löwgren

DS 91: Proceedings of NordDesign 2018, 2018.

Appended papers have been removed from the electronic thesis.

Available from: <https://www.designsociety.org/publication/40939/>

Paper 5

Verification Staircase: a Design Strategy for Actionable Explanations.

M. Lindvall and J. Molin

In proceedings of the Workshop on Explainable Smart Systems for Algorithmic Transparency in Emerging Technologies, 2020.

Appended papers have been removed from the electronic thesis.

Available from: <http://ceur-ws.org/Vol-2582/paper5.pdf>

Paper 6

Designing for the Long-Tail of Machine Learning.

M. Lindvall and J. Molin

In Workshop on Emerging Perspectives in Human-Centered Machine Learning at CHI '19, 2019.

The paper was reformatted, but is otherwise unchanged.

Appended papers have been removed from the electronic thesis.

Available from: [arXiv:2001.07455](https://arxiv.org/abs/2001.07455)

Paper 7

Proactive Construction of an Annotated Imaging Database for Artificial Intelligence Training.

C. B. Stadler, M. Lindvall, C. Lundström, A. Bodén, K. Lindman, J. Rose, D. Treanor, J. Blomma, K. Stacke, N. Pinchaud, M. Hedlund, F. Landgren, M. Woisetschläger, and D. Forsberg

Journal of Digital Imaging. 34, 105-115, 2021.

Appended papers have been removed from the electronic thesis.

Available from: doi:10.1007/s10278-020-00384-4

FACULTY OF SCIENCE AND ENGINEERING

Linköping Studies in Science and Technology, Dissertations No. 2157, 2021
Division of Media and Information Technology

Linköping University
SE-601 74 Norrköping, Sweden

www.liu.se