TOWARDS INTERPRETABLE MACHINE LEARNING IN MEDICAL IMAGE ANALYSIS

by
Haomin Chen

A dissertation submitted to The Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy

Baltimore, Maryland
January, 2023

© 2023 Haomin Chen
All rights reserved

Abstract

Over the past few years, Machine Learning (ML) has demonstrated human expert level performance in many medical image analysis tasks. However, due to the black-box nature of classic deep ML models, translating these models from the bench to the bedside to support the corresponding stakeholders in the desired tasks brings substantial challenges. One solution is interpretable ML, which attempts to reveal the working mechanisms of complex models. From a human-centered design perspective, interpretability is not a property of the ML model but an affordance, i.e., a relationship between algorithm and user. Thus, prototyping and user evaluations are critical to attaining solutions that afford interpretability. Following human-centered design principles in highly specialized and high stakes domains, such as medical image analysis, is challenging due to the limited access to end users. This dilemma is further exacerbated by the high knowledge imbalance between ML designers and end users. To overcome the predicament, we first define 4 levels of clinical evidence that can be used to justify the interpretability to design ML models. We state that designing ML models with 2 levels of clinical evidence: 1) commonly used clinical evidence, such as clinical guidelines, and 2) iteratively developed clinical evidence with end users are more likely to design models that are indeed interpretable to end users.

In this dissertation, we first address how to design interpretable ML in medical image analysis that affords interpretability with these two different levels of clinical evidence. We further highly recommend formative user research as the first step of the interpretable model design to understand user needs and domain requirements.

We also indicate the importance of empirical user evaluation to support transparent ML design choices to facilitate the adoption of human-centered design principles. All these aspects in this dissertation increase the likelihood that the algorithms afford interpretability and enable stakeholders to capitalize on the benefits of interpretable ML. In detail, we first propose neural symbolic reasoning to implement public clinical evidence into the designed models for various routinely performed clinical tasks. We utilize the routinely applied clinical taxonomy for abnormality classification in chest x-rays. We also establish a spleen injury grading system by strictly following the clinical guidelines for symbolic reasoning with the detected and segmented salient clinical features. Then, we propose the entire interpretable pipeline for Uveal Melanoma (UM) prognostication with cytopathology images. We first perform formative user research and found that pathologists believe cell composition is informative for UM prognostication. Thus, we build a model to analyze cell composition directly. Finally, we conduct a comprehensive user study to assess the human factors of human-machine teaming with the designed model, e.g., whether the proposed model indeed affords interpretability to pathologists. The human-centered design process is proven to be truly interpretable to pathologists for UM prognostication. All in all, this dissertation introduces a comprehensive human-centered design for interpretable ML solutions in medical image analysis that affords interpretability to end users.

Thesis Readers

Dr. Mathias Unberath (Primary Advisor)
Assistant Professor
Department of Computer Science
Johns Hopkins University

Dr. Gregory D. Hager
Mandell Bellmore Professor
Department of Computer Science
Johns Hopkins University

Dr. Chien-Ming Huang
Assistant Professor
Department of Computer Science
Johns Hopkins University

Acknowledgements

First and foremost, I would like to express my sincere appreciation to my advisor Prof. Mathias Unberath for his genuine patience, tremendous support, and insightful guidance throughout my Ph.D. career. His insightful criticism and suggestions on how to improve my work have impacted me far more than any document could possibly reveal. He encouraged me to see beyond projects and simply marvel at the wonderful future of daily life with our work, and many times he lifted my head out of the technical details and helped me see the significance of my contribution to the entire research community. One of his golden sentences that most impressed me is *"How to formalize research ideas that can be presented on New York Times?"* I will never forget that research is not only about technical tricks limited to specific problems, but more importantly, innovative ideas and products that can truly improve human life.

Next, I would wholeheartedly appreciate my second advisor Prof. Gregory D. Hager for his fundamental advice and suggestions on how to do research in computer vision. He revealed a flamboyant world to me with numerous insightful research ideas. I am hugely impressed by his quick speed to accept new ideas and knowledge and his innovative feedback on these ideas. I learned not only basic but also innovative concepts during the brainstorming meetings. I am also very thankful to Prof. Chien-Ming Huang for being my committee member. I especially acknowledge Prof. Chien-Ming for his interest in my research and for his helpful instructions.

Thirdly I would thank Prof. Alvin Liu for serving as my GBO committee member. My Ph.D. career is highly dedicated to the uveal melanoma prognostication project using deep learning techniques, founded by the Emerson Collective Cancer Research Fund. In this project, Alvin prepared one of the largest uveal melanoma cytopathology dataset in the world, which is time-consuming and labor-intensive. The follow-up survival status further makes our proposed system more meaningful in clinical practices. Besides, I also want to thank Prof. Zelia M. Correa who is involved in this interdisciplinary project with potentially unprecedented clinical impact.

Besides, I would like to thank Prof. Rama Chellappa, Prof. Archana Venkatraman, and Prof. Vladimir Braverman for taking their valuable time in participating in my Ph.D. GBO exam and providing helpful suggestions and comments on my coursework and research study.

I am more than grateful to be advised by Dr. Le Lu in PAII and NVIDIA, where I spent multiple semesters very worthwhile as an intern. It was from him that I realized that a computer science researcher in the medical image field should think much more beyond the computer science perspective. Only after we bear in mind that "Doctors and patients first" can we serve them well. In PAII and NVIDIA, I also learned so much from many professional experts. I made amazing friends here, *i.e.*, Shun, Adam, Daguang, Zhuotun, Yingda, Fengze, Chaochao, Can, Bowen, Weijian, JingZheng, Dakai, Dazhou, Ke, Ling, Kang, Yuankai, Yuhang, Ashwin, and our administrative staff Yizhi.

Last but not least, during my Ph.D. study, I am fortunate to meet and/or work with so many nice and professional postdocs, Ph.D. students, and visiting students in ArcadeLab, *i.e.*, Catalina Gomez, Cong Gao, Max Li, Wenhao Gu, Roger Soberanis, Nathan Drenkow, Benjamin D. Killeen, Xingtong Liu, Jasmine Cho, Hao Ding, Baichuan Jiang, Jonas Winter, Jieying Wu, Anna Zapaischikova, Kinjal Shah, Yiqing Shen, Mareike Thies, Philipp Nikutta and those in CIRL Lab, *i.e.* Molly O'Brien, Andrew Hundt, Jonathan Jones, Tae Soo Kim, Michael Peven, Jin Bai, Weiyao Wang, and Yotam Barnoy. Not only do we discuss research with each other, but also we play and eat together. Those happy memories won't be forgotten.

To my beloved parents and family for their unconditional and unreserved support.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

In the last decade, traditional Machine Learning (ML) is quickly developed to assist humans in simple tasks, such as e-Mail filtering [4], digit recognition [5], and detection of oil spills in satellite radar images [6]. Since the emergence of AlexNet [7] around 10 years ago, deep MLs has become a game-changing technique to achieve human-comparable performance for various tasks, such as natural image reasoning and medical image analysis. Although these tasks are much more complicated compared to those solved by traditional ML, the ultimate goals of these ML models remain the same as those of traditional ML models, which is to assist the end users for specific tasks rather than replacing them. However, due to the black-box nature of classic deep ML models, translating these models from the bench to the bedside to support the corresponding stakeholders in the desired tasks brings substantial challenges. When stakeholders interact with ML tools to reach decisions, they may be persuaded to follow ML's recommendations that may be incorrect or promote unintended biases against vulnerable populations, all of which can have dreadful consequences [8]. These circumstances motivate the need for trustworthy ML systems and have sparked efforts to specify the different requirements that ML algorithms should fulfill. Compared to other imaging problems, trustworthy ML models is much more desirable in medical image analysis because of the high stakes involved in most decisions that impact human lives. Most of these recent efforts focus on achieving a certain on-task performance requirement but neglect that for assisted decision making not ML system performance alone, but human-ML team performance is the most pertinent to patient outcome. How to achieve adequate human-machine teaming performance, however, is debated. While some argue that rigorous algorithmic validation, e. g., similar to the evaluation of drugs, tests, or devices, demonstrates safe and reliable operation and may thus be sufficient for successful human-machine teaming [9, 10], others reason that interpretability in an ML model, e. g., by revealing its working mechanisms and presenting a proper interface, is necessary to invoke user trust and achieve the desired human-machine teaming performance [11–13]. The inability to make the decision making process interpretable might affect the misuse and disuse of ML models in the clinical domain, as the utility of the model might be limited if it does not reveal the reasoning process, limitations, and biases [14]. We believe that this dichotomy is artificial in that, first, rigorous validation and interpretability are not mutually exclusive, and second, both approaches augment an ML model with additional information in hopes to justify (in other words, make interpretable) the recommendation's validity which is hypothesized to achieve certain human-factors engineering goals such as understandability, reliability, trust and etc.

Designing ML algorithms that are interpretable is fundamentally different from merely designing ML algorithms. The desire for interpretability adds a layer of complexity that is not necessarily computational. Rather, it involves human factors, namely the users to whom the ML algorithm should be interpretable. As a consequence, the interpretability of an algorithm is not a property of the algorithm but a relationship between the interpretable ML algorithm and the user processing the information. Such relationship can be understood as an *affordance*, a concept that is commonly employed when designing effective Human-Computer Interactions (HCIs) [15], and we argue that interpretability in ML algorithms should be viewed as such. There are several consequences from this definition:

- Developing interpretable ML algorithms is not purely computational.
- Specific design choices on the mechanisms to achieve explanations or interpretations may be suitable for one user group, but not for another.
- Creating interpretable ML systems without prior groundwork to establish that it indeed affords interpretability may result in misspent effort.

Given the user- and context-dependent nature of interpretability, it is essential to understand the target audience and to validate design choices through iterative empirical user studies to ensure that design choices of interpretable models are grounded in a deep understanding of the target users and their context. In addition, to maintain a user-centered approach to design from the early stages, rapid prototyping with users provides feedback on the current, low- to high-fidelity embodiment of the system that is going to be built eventually. Involving users early by exposing them to low-fidelity prototypes that mimic final system behavior allows designers to explore multiple alternatives before committing to one pre-determined approach that may not be understandable nor of interest to end users.

However, following a human-centered design approach to build interpretable ML systems for highly specialized and high stakes domains, such as healthcare, is challenging. The barriers are diverse and include: 1) the high knowledge mismatch between ML developers and the varied stakeholders in medicine, including providers, administrators, or patients; 2) availability restrictions or ethical concerns that limit accessibility of potential target users for iterated empirical tests in simulated setups for formative research or validation; 3) challenges inherent to clinical problems, including the complex nature of medical data (e.g., unstructured or high dimensional) and decision making tasks from multiple data sources; and last but not least, 4) the lack of ML designers' training in design thinking and human factors engineering.

Furthermore, there are multiple "interpretability" techniques and choices, such as the interpretable working mechanism or user-friendly HCI. Simply selecting a "interpretability" technique, without incorporating and consulting target users puts the resulting ML models at risk of not achieving the desired interpretability. The human-centered design approach addresses this challenge through iterative empirical studies that over time guide the development and refinement of the technical approach such that, upon completion, the design choices are well justified by empirical target user feedback. This approach may not always be feasible in healthcare due to accessibility and availability barriers of target users. To address this limitation while still enabling technological progress in interpretable ML, we introduce four distinct levels of evidence. These levels allow designers to classify the level of confidence one may have that the specific design choices will indeed result in a model that affords interpretability. The levels of evidence are based on increasingly thorough approaches to understanding the chosen end users in the context of the envisioned task [16]:

- **Level 0: No evidence.** No dedicated investigations about the end users are performed to develop interpretable ML systems.

- **Level 1: One-way evidence.** Formative user research techniques, such as surveys and diary studies, are only performed once without further feedback from end users about the findings extracted from the research phase, resulting in one-way evidence. Such user research suffers risks of potential bias in concluding about justification of interpretability because there is no opportunity for dialog, i.e., designers may ask irrelevant questions or target users may provide non-insightful, potentially biased responses.

- **Level 2: Public evidence.** Public evidence refers to information about target user knowledge, preference, or behavior that is public domain and vetted in a sensible way. Public evidence includes clinical best practice guidelines, Delphi consensus reports, peer-reviewed empirical studies of closely related approaches in large cohorts, or well documented socio-behavioral phenomena.

- **Level 3: Iteratively developed evidence.** Iteratively developed evidence is interpretability evidence that is iteratively refined through user feedback where designers and end users communicate with each other throughout method development. The purpose of iteratively validating and refining the current interpretability mechanism is to identify any potential bias in the assumptions that motivate the interpretability technique while ensuring that it is understandable to end users.

Being actively cognizant of the level of evidence that supports the development enables trading off development efforts between ML method development vs. gathering richer evidence in support of the intended developments.

Starting from the considerations around designing and validating interpretable ML for healthcare presented above, we aim to actively consider and work closely with the end users during the design, construction, and validation of ML models for medical imaging problems. Acknowledging the barriers to widespread adoption of human-centered design techniques to develop interpretable ML in healthcare, we highlight the need to ground and justify design choices in a solid understanding of the users and their context when adding interpretability or other human factors-based goals to ML systems for medical image analysis. By raising awareness of the user- and context-dependent nature of interpretability, we consider a trade-off between efforts to 1) better ground their approaches on user needs and domain requirements and 2) commit to technological development and validation of possibly interpretable systems. In this way, we increase the likelihood for algorithms that advance to the technological development stage to afford interpretability, because they are well grounded and justified in user and context understanding. This may mitigate misspent efforts in developing complex systems without prior formative user research, and help us make accurate claims about interpretability and other human factors engineering goals when building and validating the model.

In this dissertation, we develop task-specific interpretable ML for various medical image analysis problems with high-stakes decision makings by actively considering and working closely with the end users during the design, construction, and validation of ML models for medical imaging problems. This approach increases the likelihood for algorithms that advance to the technological development stage to afford interpretability.

As described in the previous introduction, the human-centered design approach performs iterative empirical studies that over time guide the development and refinement of the technical approach such that, the design choices are well justified by empirical target user feedback. In addition, the included medical image analysis tasks in this dissertation are all high-stakes problems. Potential bias in concluding the clinical evidence may result in the mismatch between the justification of interpretability and the context of the end users. Therefore, the end users may be unwilling to use the developed models or the models are not understandable or useful to the end users. Thus, only public evidence that the end users commonly use; and iteratively developed evidence that is iteratively refined through user feedback are suitable to be the justification of interpretability to design the models. In one aspect, we develop interpretable models with public evidence, such as clinical guidelines in clinical routine practices for clinical experts. In the other aspect, for clinical problems that are not routinely performed, such as tasks beyond end users' ability and knowledge, we perform formative user research to iteratively develop the clinical evidence to design the interpretable model. We further conduct user studies to assess the interpretability and human factors of the designed models.

1.1 Challenges and Our Contributions

We introduce two works of building interpretable ML models for routinely performed clinical problems with the widely-used clinical guidelines and one work of building interpretable ML models for a medical image analysis task that is beyond clinical experts' ability with iteratively developed clinical evidence. For the first two works, we develop neural-symbolic models with implemented public evidence for each medical image analysis task, as presented in Chapter 3 to Chapter 4. On one hand, neural reasoning with deep learning is more robust with noisy and ambiguous data but lacks interpretability. On the other hand, symbolic reasoning can naturally leverage symbolic representations of clinical knowledge such as clinical guidelines but is intolerant of ambiguous and noisy data [17]. Combining the two reasoning methods and developing neural-symbolic models makes it easy to implement clinical knowledge into the entire deep learning system and robust to noisy and ambiguous data. In our design process, deep Convolutional Neural Network (CNN) architectures are implemented to extract robust and meaningful visual features. The symbolic reasoning modules with implemented clinical knowledge further analyze the extracted visual features which have the potential to afford interpretability to end users. The implemented clinical knowledge is the relevant knowledge that is understandable and useful for the end users, which is also known as the justification of interpretability in the designed ML. Before the model design, we guarantee the justification of interpretability to be relevant knowledge by the public clinical evidence that is publicly agreed on in clinical publications. These two works are:

1.1.1 Abnormality Classification in Chest X-Rays with Clinical Guideline: Clinical Taxonomy

With the rapid development of deep CNNs and their success in many computer vision tasks [7], Chest X-Ray (CXR) computer-aided diagnosis (CAD) has received considerable research attention [18–20]. These efforts have met success and typically approach the problem as a standard multi-label classification scenario, which attempts to make a set of individual binary predictions for each disease pattern under consideration. Yet, organizing diagnoses or observations into ontologies and/or taxonomies is crucial within radiology, e.g., RadLex [21], with CXR interpretation being no exception [22, 23]. This importance should also be reflected within CAD systems. For instance, when uncertain about fine-level predictions, e.g., *nodules* vs. *masses*, a CAD system should still be able to provide meaningful parent-level predictions, e.g., *pulmonary nodules and masses*. This parent prediction may be all the clinician is interested in anyway. Moreover, elegantly addressing the problem of incompletely labeled data is another benefit of incorporating taxonomy. Radiologists may only report general/coarse disease labels because of the imaging conditions or because the diseases are unrelated to the purpose of patient admission. For example, imaging conditions may have only allowed a radiologist to report "opacity", instead of a more specific observation of "infiltration" vs. "atelectasis". As a result, it is clinically beneficial for CAD systems to not only report fine-grained labels but also report labels higher up in the clinical taxonomy. For these reasons, we present a deep hierarchical multi-label classification (HMLC) approach for CXR CAD with a label taxonomy constructed with the reference of clinical taxonomy [21].

To the best of our knowledge, we are the first to outline an HMLC CAD system and implement clinical taxonomy for medical imaging and the first to characterize performance when faced with incompletely labeled data. Our straightforward, but effective, HMLC approach results in the highest mean Area Under Curve (AUC) value yet reported for the Prostate Lung Colorectal and Ovarian (PLCO) dataset. In incompletely labeled data scenarios, HMLC can garner even greater boosts in classification performance. The method also has a high potential to be interpretable to radiologists because of following the widely used clinical taxonomy. The methods are detailed in Chapter 3, and were presented at one conference and in a journal article:

[24] Chen, H., Miao, S., Xu, D., Hager, G.D. and Harrison, A.P., 2019, May. Deep hierarchical multi-label classification of chest X-ray images. In International conference on medical imaging with deep learning (MIDL) (pp. 109-120). PMLR

[25] Chen, H., Miao, S., Xu, D., Hager, G.D. and Harrison, A.P., 2020. Deep hierarchical multi-label classification applied to chest X-ray abnormality taxonomies. Medical image analysis, 66, p.101811.

1.1.2 Splenic Injury Grading in Whole Body CT Scans with Clinical Guideline: AAST Grading Criterion

Splenic injury is the most common solid organ injury in adult blunt abdominal trauma [26, 27]. In 2018, the American Association for the Surgery of Trauma (AAST) Patient Assessment Committee (PAC) introduced an updated AAST splenic organ injury scale (OIS) for treatment decision-making based on admission abdominopelvic CT examination. Treatment options vary by injury severity and include routine observation for low grade injuries, and urgent angioembolization or splenectomy to control hemorrhage in high grade injuries.

In a survey of AAST member practices in the management of blunt splenic injury, only 45% of respondents reported routine use of the AAST splenic OIS for blunt splenic trauma by radiologists [28]. Even assuming the ideal circumstance of ubiquitous adoption and reporting, classification systems are prone to variability in the perceived grades among readers with varying levels of experience, and variable subspecialization. Recent preliminary data from a retrospective multicenter multileader American Society of Emergency Radiology study on blunt splenic trauma indicates only moderate agreement under research conditions. In practice, radiologists are subject to shifting circumstances in their clinical environment with respect to study volume, reading room distractions, and fatigue-related performance degradation, such as from circadian rhythm disruptions after multiple consecutive night shifts [29–33]. Furthermore, clinical decision making in this high-stakes setting must be rapid, as the spleen is a highly vascular organ and severe injury carries the risk of multi-organ system failure and death from exsanguination [34]. However, admission trauma CT interpretation is time-consuming. Among expert trauma radiologists, interpretation turnaround times for severely injured patients commonly exceed 20-30 minutes [29].

Automated AAST grading could potentially provide an objective, accurate, second-reader capability that grants users agency in addressing disagreement between the automated method and their own domain expertise [35, 36]. More importantly, the automated AAST grading can be considered as a guided read. Domain expertise can double check whether findings in automated systems truly exist or not to confirm or not trust the automated AAST grade diagnosis for further treatment, such as mobilization of vascular surgeon. To this end, we leverage interpretable deep learning approaches and expert knowledge to develop a novel automated method with a hierarchical rule-based system that follows the AAST grading pipeline and predicts the AAST splenic OIS using the most salient CNN-extracted features of the grading system, namely active bleeding, pseudoaneurysm, and splenic parenchymal disruption [37]. The methods are detailed in Chapter 4, and were presented in a journal article:

[38] Chen, H., Unberath, M. and Dreizin, D., 2022. Toward automated interpretable AAST grading for blunt splenic injury. Emergency Radiology, pp.1-10.

We also build an interpretable ML models for tasks that clinical experts do not have the ability to accomplish with current clinical knowledge. We iteratively develop the clinical evidence to be used in the model by formative user research with user feedback. We additionally conduct a user study to assess the interpretability and the human factors of our models to the clinical experts. The clinical scenario and the ML solution are described below:

1.1.3 Uveal Melanoma Cancer Subtyping with Cytopathology Images by Analyzing What Pathologists Believe to be Salient: Cell Type Composition

Uveal Melanoma (UM) is the most common primary intraocular malignancy in adults [39]. According to a recent study, there exist two subtypes in UM that can be identified based on its Gene Expression Profile (GEP): The first subtype exhibits low metastatic risk, while the second subtype has been linked to high metastatic risk. However, even after 10 years of development, GEP is still only available in the United States. The technique is also expensive and has a long turnaround time. There also exists unexpected clinical surprises such as early death with GEP results. A more accessible test for UM subtyping is, therefore, highly desirable.

In addition to GEP, microscopic Cytology of Fine Needle Aspirates images is also created from the biopsy. There is increasing evidence that there exist imaging-derived biomarkers that are informative for prognosis [40]. In the particular case of UM prognostication, there is huge potential in using imaging-derived biomarkers to determine GEP subtype and metastatic risk directly from cytopathology slides. Pathologists also believe the imaging-derived biomarkers are hidden in the overall cell appearance composition in the cytopathology images. While it is impossible even for highly trained pathologists to derive this information from cytopathology images, learning-based algorithms that discover associations between intensity patterns in cytology images and GEP subtype are promising [41, 42]. However, as “black box” models that perform a super-human task, these algorithms do not offer insights beyond the final recommendation to the human decision makers, which has been linked to automation bias and over-trust or dis-trust in such systems [43, 44]. A more interpretable algorithm design may enable humans to better calibrate their trust in the recommendation, which would be an important feat for high-stakes decision making.

To reach this goal, we need to facilitate or even automate quantitative analysis of cytopathology images. To this end, we first develop an interactive tool to extract high-quality image regions from cytopathology images. Our envision will be beneficial in two ways: First, it can be deployed in pathologist-centric workflows to guide pathologist review, thereby reducing the experts' workload. Second, the tool provides an opportunity for pathologists to guide algorithmic evaluation, e. g. by refining the content that is submitted for the following automated analysis of the slide, e. g. for GEP classification. Such an interactive design may prove beneficial in building trust, accelerating workflows, and reducing mistakes, of both automated algorithms and pathologists.

After interactively extracting high-quality image regions, we develop an automatic system for interpretable UM subtype classification from cytology images. Before designing the models, we first perform formative user research to understand the need and knowledge of the pathologists. They believe cell composition is informative for UM prognostication. The designed method is based on the idea from the formative user research that biopsy samples of the two UM subtypes should differ in overall cell composition. Thus, we propose an algorithm that enables high level, rule-based symbolic reasoning on the overall cell composition of the cytopathology images extracted by deep CNN, which would be interpretable and could easily be verified by human users such as pathologists.

Finally, we conducted a comprehensive user study with 4 trained pathologists to test the affordance of interpretability in the proposed system. In the user study, pathologists make diagnoses for GEP classes from cytopathology images, which is a task beyond current clinical knowledge, with or without AI assistance. Human factors are analyzed and indicate that the designed cancer subtyping model for UM is truly interpretable to end users. The methods are detailed in Chapter 5 and Chapter 6, and were presented at two conferences. Moreover, we are currently preparing a journal article that summarizes the results presented in Chapter 6.

[42] Chen, H., Liu, T.Y., Correa, Z. and Unberath, M., 2020, October. An Interactive Approach to Region of Interest Selection in Cytologic Analysis of Uveal Melanoma Based on Unsupervised Clustering. In International Workshop on Ophthalmic Medical Image Analysis (pp. 114-124). Springer, Cham.

[45] Chen, H., Liu, T.Y., Gomez, C., Correa, Z. and Unberath, M., 2021. An interpretable Algorithm for uveal melanoma subtyping from whole slide cytology images. arXiv preprint arXiv:2108.06246.

1.1.4 Other Contributions to Interpretable Machine Learning in Medical Image Analysis

A series of contributions to the current state of interpretable machine learning in medical image analysis have been made, that are in close connection to this thesis. The main developments include:

- A systematic review of interpretable ML in medical image analysis together with guidelines that recommend the human-centered design. Published in NPJ Digital Medicine [16].

- An interpretable pelvic fracture detection in pelvic x-rays by comparing anatomical vertical asymmetry. Published in European Conference in Computer Vision (ECCV) [46].

- Attention-based cancer prognostication for uveal melanoma with t-SNE clustering visualization. Published in Ophthalmology Science [47] and SSRN Electronic Journal [48].

- Gene expression profile prediction for uveal melanoma with image regions. Published in Ophthalmology Retina [41].

- Survival prediction for uveal melanoma with cell feature analysis. We are currently submitting the paper.

- Causal inference for pelvic fracture tile grade. We are currently submitting the paper.

I also have contributed to semantic and realistic style transfer from 2D style images to 3D scenes. The paper is currently submitting.

1.2 Dissertation Statement

In this dissertation, we emphasize the importance of and propose the approaches to building interpretable ML for medical image analysis incorporating end users in a human-centered design. We indicate that interpretability is not a property, but an affordance of interpretable ML systems, i.e., a relationship between models and end users. Efforts to build ML systems that afford interpretability in the healthcare context should go beyond computational advances, which is not common practice in the context of interpretable ML for medical image analysis. Considering the wide gap between clinical end users and ML designers, we propose formative user research to understand the context of end users which greatly reduces the risks of building models that are not actually interpretable to end users. We also emphasize the importance of empirical user testing to assess the interpretability of the built models and further avoid unexpected shortcomings of the designed models in real clinical practices. In addition, we proposed neural-symbolic reasoning models to implement selected clinical knowledge with neural networks for both routinely performed clinical tasks and tasks that are beyond current end users' ability. The entire design procedure of formative user research, neural-symbolic reasoning, and empirical user testing achieves a comprehensive understanding of the user contexts, ensures the model interpretability to end users, and enables iterative refinement of user-friendly ML models for real clinical practices.

1.3 Overview

The overview of this dissertation is illustrated as the following.

In Chapter 1 (this chapter), we introduce the need for interpretable ML in medical image analysis. We discuss the underlying challenges and our contributions to this dissertation topic.

In chapter 2, we summarize previous works on interpretable ML in medical image analysis.

In Chapter 3, we propose a hierarchical multi-label classification for CXRs with clinical guideline: clinical taxonomy.

In Chapter 4, we propose an automatic AAST grading technique for splenic injury by following AAST clinical guidelines.

In Chapter 5, we propose an automatic but interactive high-quality region extraction algorithm for UM cytopathology images.

In Chapter 6, we propose an entire pipeline for the interpretable rule-based algorithm to predict GEP classes from UM cytopathology images by analyzing cell compositions, which is believed to be salient features by pathologists.

In Chapter 7, we summarize and conclude this dissertation.

Chapter 2

Related Work

In this chapter, we discuss related works in literature in the scope of convolutional neural networks, deep learning in medical image analysis, and interpretable Artificial Intelligence (AI) in medical image analysis. More detailed related work about interpretable deep learning in medical image analysis can be found in my co-first-authored systematic review paper:

> Haomin Chen, et al. "Explainable medical imaging AI needs a human-centered design: guidelines and evidence from a systematic review." NPJ digital medicine 5.1 (2022): 1-15. [16]

Related work about each challenging clinical tasks mentioned in the introduction chapter is included in Chapter 3 to Chapter 6.

2.1 Interpretable Convolutional Neural Network

CNN is one type of machine learning model for processing data that has a grid pattern, such as images, which is inspired by the organization of animal visual cortex [49, 50], and designed to automatically and adaptively learn spatial hierarchies of features, from low- to high-level patterns. Different from most recent radiomics studies which use hand-crafted feature extraction techniques, such as texture analysis, followed by conventional machine learning classifiers, such as random forests and support vector machines [51, 52], CNN extracts features automatically with the combination of convolutional, pooling and fully connected layers. The complex combination of these kernel operations makes CNN easy to automatically detect significant features without any human supervision which made it the most used machine learning technique these years. CNN has already been proven to reach human expert level performance in various tasks, such as face recognition [53], lung nodule detection [54] and natural image classification [55]. However, such overwhelming performance is at the price of not being explainable and CNN is also described as a black box model. It is difficult for humans to understand the reason behind CNN making a certain decision.

Research on interpretable CNN has gained momentum to provide knowledge and insights into neural networks. There are two types of CNN self-reasoning: Explainability (or model-agnostic) and Interpretability (or model-specific) [56]. Explainability is the rationale behind the decision made by CNN[57]. The CNN remains to be a black box model, but an ad-hoc model is applied to explain the output of the CNN. The ad-hoc model is typically a simple self-reasoning model. Pixel-attribution model such as saliency map is the most commonly applied ad-hoc model [58, 59]. The visualization of pixel-attribution highlights regions used for CNN outputs. Besides, region attribution is also calculated with image region occlusion in LIME [60]. In addition, Shapley values are used to evaluate the feature attribution [61]. In contrast, interpretability is intrinsic, meaning that the model structure of CNN self-explains the functioning. Attention mechanism [62] weights feature by pixel-level importance inside of the network structure. Region proposal learning with natural images intuitively implements prototypes for image classification reasoning [63]. Intrinsically interpretable models are usually constructed based on the defined justification of interpretability, which varies in different tasks.

There also exist two scales of explanation that are agreed on within the research community: local explanation and global explanation. Local methods explain the individual predictions of ML algorithms. Local explanation approaches have currently received much attention. Popular local explanable methods are Shapley values [61, 64] and counterfactual explanations [65–67]. Counterfactual explanations rely on what-if scenarios to explain the model predictions [68]. According to the current literature in the social sciences [69], counterfactual explanations are contrastive explanations and are supported by a few reasons, so they are "good" explanations. Another local explanation approach: the Shapley values explain how the deep CNN features collaborate together to generate the final prediction [61, 64]. In contrast, the global explanation aims to explain the entire model behavior, i.e., how the model behaves in general with all samples of the dataset, instead of the model behavior of every specific sample. Two main approaches for global explanation are the model's feature importance and feature effect. Feature importance assesses the CNN features' relevancy to the model predictions. Permutation feature analysis [70, 71] is a widely-used importance metric. Some permutation feature analysis removes the features from the training data and retrains the CNN models to test the importance of the removed feature [72]. Another permutation feature analysis approach is based on variance measures [73].

2.2 Interpretable Deep Learning in Medical Image Analysis

Deep learning methods such as CNN have proven to achieve human expert level performance in multiple medical image analysis tasks such as CXR disease classification [74], emphysema quantification in Computed Tomography (CT) scans [75], placenta segmentation in ultrasound imaging [76] and etc. However, most of these models are black box models and clinical stakeholders such as clinical experts and patients cannot understand why black box models make certain decisions. Moreover, medical images are commonly related to high-stakes decision makings impacting lives. Clinical stakeholders cannot afford to blindly trust or distrust black box models to guide follow-up treatment because all deep learning models may have unexpected mistakes.

Thus, there is a growing interest in developing interpretable CNN models for medical image analysis. We performed a systematic review of interpretable ML in medical image analysis [16]. We follow the PRISMA pipeline [77] for the systematic review and 68 articles were included for information extraction. Details of the PRISMA process can be found in the systematic review paper. In the systematic review, we group additional considerations about human factors and clinical context into six themes according to the initial review, iteratively defined prior to data extraction and abbreviated to *INTRPRT*; the themes are incorporation (IN), interpretability (IN), target (T), reporting (R), prior (PR), and task (T). *Incorporation* refers to the communication and cooperation between designers and end users before and during the construction of the transparent model. Formative user research is one possible strategy that can help designers to understand end users' needs and background knowledge [35, 78], but other approaches exist [79]. *Interpretability* considers the technicalities of algorithmic realization of a transparent ML system. *Target* determines the end users of the transparent ML algorithms. *Reporting* summarizes all aspects pertaining to the validation of transparent algorithms. This includes task performance evaluation as well as the assessment of technical correctness and human factors of the proposed transparency technique (e.g., intelligibility of the model output, trust, or reliability). *Prior* refers to previously published, otherwise public, or empirically established sources of information about target users and their context. This prior evidence can be used to conceptualize and justify design choices around achieving transparency. Finally, *task* specifies the considered medical image analysis task, such as prediction, segmentation, or super resolution, and thus determines the clinical requirements on performance. These themes should not be considered in isolation because they interact with and are relevant to each other. For example, the technical feasibility of innovative transparency mechanisms based on the desired task may influence both, the priors that will be considered during development as well as the incorporation of target users to identify and validate alternatives.

We structured the findings in the systematic review using the defined six themes of the *INTRPRT*, the adequacy of which was confirmed during data extraction.

IN: Incorporation

A common trend among included studies (n=33) was that the presented methods were developed by multidisciplinary clinician-engineering teams, as was evidenced by the incorporation of clinical specialists, such as physicians, radiologists, or pathologists, in the study team and on the author lists. In light of the current bias towards clinicians as end users of transparent ML algorithms, this observation suggests that designers may have communicated with a limited subset of the intended end users. However, no formative user research is explicitly described or introduced in these articles to systematically understand the end users before implementing the model. Further, incorporating clinical experts did not have a considerable impact on whether clinical priors or standard or care guidelines (i.e., Level 2 evidence) were used to build the ML system (39%/44% articles with/without the incorporation of end users use clinical priors). Regarding the technical approach to provide transparency, the incorporation of medical experts motivated designers to incorporate prior knowledge directly into the model structure and/or inference for medical imaging (73%/64% articles with/without the incorporation of end users do not need a second model to generate transparency).

IN: Interpretability

Transparency of ML systems was achieved through various techniques, including attention mechanisms (n=15), use of human-understandable features (n=11), a combination of deep neural networks and transparent traditional ML methods (n=7), visualization approaches (n=5), clustering methods (n=4), uncertainty estimation

/ confidence calibration (n=3), relation analysis between outputs and hand-crafted features (n=3), and other custom techniques (n=20).

The use of an attention mechanism was the most common technique for adding transparency. Attention mechanisms enabled the generation of pixel-attribution methods [80] to visualize pixel-level importance for a specific class of interest [81–95]. In segmentation tasks, where clinically relevant abnormalities and organs are usually of small sizes, features from different resolution levels were aggregated to compute attention and generate more accurate outcomes, as demonstrated in multiple applications, e.g., multi-class segmentation in fetal Magnetic Resonance Imagings (MRIs) [90] and multiple sclerosis segmentation in MRIs [93]. Clinical prior knowledge was also inserted into the attention mechanism to make the whole system more transparent. For instance, [83] split brain MRIs into 96 clinically important regions and used a genetic algorithm to calculate the importance of each region to evaluate Alzheimer's Disease (AD).

Human-understandable features, e.g., hand-crafted low-dimensional features or clinical variables (age, gender, etc.) were frequently used to establish transparent systems. There existed two main ways to use human-understandable features in medical imaging: 1) Extracting hand-crafted features, e.g., morphological and radiomic features, from predicted segmentation masks generated by a non-transparent model [96–105] followed by analysis of those hand crafted features using a separate classification module; 2) Directly predicting human-understandable features together with the main classification and detection tasks [106–110]. In these approaches, all tasks usually shared the same network architecture and parameter weights.

Instead of explicitly extracting or predicting human-understandable features, other articles further analyzed deep encoded features with human-understandable techniques by following clinical knowledge. Techniques such as decision trees were constructed based on clinical taxonomy for hierarchical learning [103, 111–116]. Rule-based algorithms [74] and regression methods [117] were used to promote transparency of the prediction. [118] created a Graphical Convolution Network (GCN) based on clinical knowledge to model the correlations among colposcopic images captured around five key time slots during a visual examination.

We also identified various other methods to create transparent systems. These methods can be categorized as visualization-based, feature-based, region importance-based, and architecture modification-based methods. Each approach is discussed in detail below.

Visualization-based methods provide easy-to-understand illustrations by overlaying the original images with additional visual layouts generated from transparency techniques. There existed two main visualization-based methods: 1) Visualizing pixel-attribution maps: These maps may be generated using gradient-based importance analysis [119, 120], pixel-level predicted probability [121], or a combination of different levels of feature maps [122, 123]. 2) Latent feature evolution: Encoded features were evolved according to the gradient ascent direction so that the decoded image (e.g., generated with an auto-encoder technique [124]) gradually change from one class to another [125, 126].

Feature-based methods directly analyze encoded features in an attempt to make the models transparent. Various feature-based transparency method were proposed for transparent learning. [127–129] first encoded images to deep features and then clustered samples based on these deep features for prediction or image grouping tasks. Feature importance was also well-studied to identify features that are most relevant for a specific class by feature perturbation [130, 131] and gradients [132]. [133] identified and removed features with less importance for final prediction through feature ranking.

As an alternative to measure feature contribution, input region importance was also analyzed to reveal sub-region relevance to each prediction class. Image occlusion with blank sub-regions [134–136] and healthy-looking sub-regions [137] was used to find the most informative and relevant sub-regions for classification and detection tasks.

Other approaches modified the network architecture according to relevant clinical knowledge to make the whole system transparent. [90] pruned the architecture according to the degree of scale invariance at each layer in the network. [138] created ten branches with shared weights for ten ultrasound images to mimic the clinical workflow of liver fibrosis stage prediction. [139] aggregated information from all three views of mammograms and used traditional methods to detect nipple and muscle direction, which was followed by a grid alignment according to the nipple and muscle direction for left and right breasts. [140] proposed to learn representations of the underlying anatomy with a convolutional auto-encoder by mapping the predicted and ground truth segmentation maps to a low dimensional representation to regularize the training objective of the segmentation network.

Some other methods used the training image distribution to achieve transparency in classification. [141] used similar-looking images (nearest training images in feature space) to classify testing images with majority votes. Causal inference with plug-in clinical prior knowledge also introduced transparency directly to automatic systems [142–144]. Confidence calibration and uncertainty estimation methods were also used to generate additional confidence information for end users [145–147].

T: Targets

A striking observation was that none of the selected articles aimed at building transparent systems for users other than care providers. Less than half of the articles explicitly specified clinicians as the intended end users of the system (n=30). From the remaining 38 articles, 17 articles implied that the envisioned end users would be clinicians, while the remaining 21 did not specify the envisioned target users. Articles that were more explicit about their end users were more likely to rely on clinical prior knowledge (Level 2 evidence) in model design. In total, 47% of articles that specified or implied clinicians as end users implemented clinical prior knowledge in the transparent systems while only 18% of articles without end user information use clinical prior knowledge.

R: Reporting

Evaluating different properties of a transparent algorithm besides task-related metrics, especially its performance in regards to achieving the desired human factors engineering goals, complements the assessment of the ML model's intended purpose. The quality of the transparency component is currently being evaluated through four main approaches. The first one involves metrics based on human perception, such as the mean opinion score introduced in [140] to capture two expert participants' rating of the model's outcome quality and similarity to the ground truth on a 5-point scale. Using two study participants, pathologists' feedback was also requested in [132] to assess their agreement with patch-based visualizations that display features relevant for normal and abnormal tissue. The level of agreement was not formally quantified, but reported as a qualitative description. Similarly, one study participant was involved in a qualitative assessment of explanations quality in [109, 133]. These evaluations are different from empirical user studies as they are limited to a few individuals and were mostly used to subjectively confirm the correctness of the transparent component.

The second approach attempted to quantify the quality of explanations for a specific purpose (functionally-grounded evaluation [148]). For instance, some articles evaluated the localization ability of post-hoc explanations by defining an auxiliary task, such as detection [89, 114] or segmentation [94, 111, 123, 137] of anatomical structures related to the main task. They then contrasted relevant regions identified by the model with ground truth annotations. These quantitative measures (dice score, precision, recall) allowed for further comparisons with traditional explanations methods. Similarly, [141] defined a multi-task learning framework for image classification and retrieval, evaluating retrieval precision and providing a confidence score based on the retrieved neighbors as an attempt to check the learned embedding space. Capturing relevant features consistent with human intuition was proposed in [131] by measuring the fraction of reference features recovered, which were defined according to a guideline. Overall, the evaluation of explanations through auxiliary tasks required additional manual efforts to get the necessary ground truth annotations.

Properties of the explanation itself were also quantified as their usefulness to identify risky and safe predictions at a voxel-level for the main task by thresholding on their predictive uncertainty values [147]. Other properties of explanations, such as their correctness (accuracy of rules), completeness (fraction of the training set covered) and compactness (size in bytes) were measured in [113]. A measure related to completeness was defined in [114] and aimed to capture the proportion of training images represented by the learned visual concepts, in addition to two other metrics: the inter- and intra-class diversity and the faithfulness of explanations computed by perturbing relevant patches and measuring the drop in classification confidence. Other articles followed a similar approach to validate relevant pixels or features identified with a transparent method; for example, in [82] a deletion curve was constructed by plotting the dice score vs. the percentage of pixels removed and [81] defined a recall rate when the model proposes certain number of informative channels. [120] proposed to evaluate the consistency of visualization results and the outputs of a CNN by computing the *L1* error between predicted class scores and explanation pixel-attribution maps. In summary, while the methods grouped in this theme are capable of evaluating how well a method aligns with it's intended mechanism of transparency, they fall short of capturing any human factors-related aspects of transparency design.

The third, and most common approach, involved a qualitative validation of the transparent systems (n=40) by showing pixel-attribution visualizations overlaid with the input image or rankings of feature relevance, along with narrative observations on how these visualizations may relate to the main task. These qualitative narratives might include comparisons with other visualization techniques in terms of the highlighted regions or the granularity/level of details. Furthermore, following a retrospective analysis, the consistency between the identified relevant areas/features and prior clinical knowledge in a specific task was a common discussion item in 37% of all the articles (n=25); refer to articles [83, 112, 115, 135, 142] for examples. While grounding of feature visualizations in the relevant clinical task is a commendable effort, the methods to generate the overlaid information have been criticized in regards to their fidelity and specificity [56, 149]. Further, as was the case for methods that evaluate the fidelity of transparency information, these methods do not inherently account for human factors.

Lastly, transparent systems can be directly evaluated through user studies on the target population, in which the end users interact with the developed ML system to complete a task based on a specific context. In [121], the evaluation was centered on the utility of example-based and feature-based explanations for radiologists (8 study participants) to understand the AI decision process. Users' understanding was evaluated as the accuracy to predict the AI's diagnosis for a target image and a binary judgement on whether they certify the AI for similar images (and justify using multiple-choice options). Users' agreement with the AI's predictions was measured as well. The empirical evidence suggested that explanations enabled radiologists to develop appropriate trust by making an accurate prediction and judgement of the AI's recommendations. Even though radiologists could complete the task by themselves, a comparison with the team performance was not included, nor the performance of the AI model in standalone operation. An alternative evaluation of example-based explanation usefulness was performed in [146], in which pathologists (14 study participants) determined the acceptability of a decision support tool by rating adjectives related to their perceived objectivity, details, reliability, and quality of the system. Compared to a CNN without explanations, the subjective ratings were more positive towards the explainable systems. However, neither the team (expert + AI) nor expert baseline performance was evaluated. The benefit of involving a dermatologist to complete an image grouping task was demonstrated in [127], in which domain knowledge was used to constrain updates of the algorithm's training, resulting in a better grouping performance than a fully automated method. The user evaluation only measured the task performance. These studies that explicitly involve target users to identify whether the envisioned human factors engineering goals were met stand out from the large body of work that did not consider empirical user tests. It is, however, noteworthy that even these exemplary studies are based on very small sample sizes that may not be sufficiently representative of the target users. Careful planning of the study design (including hypothesis statement, experimental design and procedure, participants, and measures) that allows to properly evaluate whether the system achieves the intended goals by adding transparency to the ML system is fundamental, especially considering the resources needed and challenges involved in conducting user testing in the healthcare domain.

Even though there were articles that assessed human factors-related properties of the transparency mechanism, a striking majority of articles did not report metrics beyond performance in the main task (n=49) or did not discuss the transparency component at all (n=9). Task performance was evaluated in the majority of the articles, 91% (n=62), and most of them contrasted the performance of the transparent systems with a non-transparent baseline (n=41). Of those, 36 works (88%) reported improved performance and 5 (12%) comparable results.

PR: Priors

We differentiate two types of priors that can be used as a source of inspiration to devise transparent ML techniques: 1) Priors based on documented knowledge, and especially clinical guidelines considering the unvaried end user specification identified above; and 2) Priors based on computer vision concepts. Most (93%) articles that incorporated clinical knowledge priors (n=28) directly implemented these priors into the model structure and/or inference, while only 68% articles with computer vision priors (n=40) provided transparency by the model itself and/or the inference procedure.

A direct way to include clinical knowledge priors was through the prediction, extraction, or use of human-understandable features. Morphological features, e.g., texture, shape and edge features were frequently considered and used to support the transparency of ML systems [96, 99, 100, 102, 103, 107, 109, 118]. Biomarkers for specific problems, e.g., end-diastolic volume (EDV) in cardiac MRI [97, 105] and mean diameter, consistency, and margin of pulmonary nodules [106] were commonly computed to establish transparency. For problems with a well-established image reporting and diagnosis systems, routinely-used clinical features, e.g., Liver Imaging Reporting and Data System (LI-RADS) features for Hepatocellular carcinoma (HCC) classification [110] or Breast Imaging Reporting and Data System (BI-RADS) for breast mass [108] suggested that the ML systems may be intuitively interpretable to experts that are already familiar with these guidelines. Human-understandable features relevant to the task domain were extracted from pathology images, e.g., area and tissue structure features [96]. Radiomic features were also computed to establish the transparency of ML systems [102, 150].

Besides human-understandable features, clinical knowledge can be used to guide the incorporation of transparency within a model. Some articles (n=11) mimicked or started from clinical guidelines and workflows to construct the ML systems [83, 101, 107, 108, 131, 138–140, 143, 144]. [101, 138, 139] followed the clinical workflow to encode multiple sources of images and fused the encoded information for the final prediction. Other works followed the specific clinical guidelines of the problems to create transparent systems. [83] split brain MRIs into 96 clinical meaningful regions as would be done in established clinical workflows and analyze all the regions separately. Some other clinical knowledge priors were also presented. [111, 112, 116, 151] established a hierarchical label structure according to clinical taxonomy for image classification. [98] leveraged the transparency from the correlation between the changes of polarization characteristics and the pathological development of cervical precancerous lesions. Clinical knowledge from human experts was used to refine an image grouping algorithm through an interactive mechanism in which experts iteratively provided inputs to the model [127].

Priors that were derived from computer vision concepts rather than the clinical workflow were usually not specific or limited to a single application. The justification of transparency with computer vision priors was more general than that with clinical knowledge priors. Image visualization-based techniques to achieve transparency were most commonly considered in image classification problems. Common ways of retrieving relevance information were: Visual relevancy through attention [81, 82, 84–86, 88–95]; region occlusion by blank areas [134, 136] or healthy-looking regions [137]; and other techniques such as supervision of activated image regions by clinically relevant areas [114, 115, 117, 119, 120, 122, 123], and image similarity [121]. Feature-based computer vision transparency priors focused on the impact of feature evolution or perturbation on the decoded output. Encoded features were evolved according to the gradient ascent direction to create the evolution of the decoded image from one class to the other [113, 125, 126]. Some articles directly analyzed the feature sensitivity to the final prediction by feature perturbation [126, 130, 135] and importance analysis [104, 132, 133], feature distribution [129, 130] or image distribution based on encoded features [128, 141]. Confidence calibration and uncertainty estimation also increased the transparency of the ML systems [145–147].

Even though attempting to identify the type of prior evidence used to justify the development of a specific algorithm in each ML system, none of the included articles formally described the process to formulate such priors to achieve transparency in the proposed system. While the use of clinical guidelines and routine workflows may provide Level 2 evidence in support of the method affording transparency if the end users are matched with those priors, relying solely on computer vision techniques may not provide the same level of justification. This is because computer vision algorithms are often developed as an analysis tool for ML developers to verify model correctness, but are not primarily designed nor evaluated for use in end user-centered interfaces. The lack of justification and formal processes to inform design choices at the early stages of model development results in substantial risk of creating transparent systems that rely on inaccurate, unintelligible, or irrelevant insights for end users. Being explicit about the assumptions and evidence available in support of the envisioned transparent ML system is paramount to build fewer but better-justified transparent ML systems that are more likely to live up to expectations in final user testing, the resources for which are heavily constrained.

T: Task

Various types of medical image analysis tasks were explored in the included articles. Most of the articles (n=57) proposed transparent ML algorithms for classification and detection problems, such as image classification and abnormality detection. Three-dimensional (3D) radiology images (n=24) and pathological images (n=15) were the most popular modalities involved in the development of transparent algorithms. The complex nature of both 3D imaging in radiology and pathological images makes image analysis tasks more time consuming than 2D image analysis that is more prevalent in other specialities, such as dermatology, which motivates transparency as an alternative to complete human image analysis to save time while retaining trustworthiness. In detail, classification problems in 3D radiological images and pathological images included abnormality detection in CT scans [84, 88, 91, 100, 102, 116, 120, 131], MRIs [83, 90, 92, 97, 104, 105, 109, 110, 123, 125, 130, 135, 137, 142], pathology images [81, 87, 89–91, 94, 96, 98, 104, 129, 132–134, 141, 146] and positron emission tomography (PET) images [95]. Mammography dominated the 2D radiology image applications [103, 107, 108, 113, 117, 119, 139, 144, 150], mainly focusing on breast cancer classification and mass detection. For other 2D radiology image applications, [121, 143] aimed at pneumonia and pneumothorax prediction from chest X-rays and [138] created a transparent model for liver fibrosis stage prediction in liver ultrasound images. Classification and detection tasks were explored in other clinical specialities, including melanoma [111] and skin lesion grade prediction [90, 112, 113] in dermatology, glaucoma detection from fundus images [86, 101, 122] and retinopathy diagnosis [136] in ophthalmology, and polyp classification from colonoscopy images in gastroenterology [114, 145].

Segmentation was another major application field (n=9). Research about transparency mainly focused on segmentation problems for brain and cardiac MRIs [82, 85, 93, 99, 115, 128, 140]. Other segmentation problems included mass segmentation in mammograms [103], cardiac segmentation in ultrasound [140], liver tumor segmentation in hepatic CT images, and skin lesion segmentation in dermatological images [90]. There also existed other applications, e.g., image grouping in dermatological images [127] and image enhancement (super resolution task) in brain MRIs [147] and cardiac MRIs [140].

Most of the application tasks were routinely performed by human experts in current clinical practice (n=60). A much smaller sample of articles (n=4) aimed to build transparent systems for much more difficult tasks where no human baseline exists, e.g., 5-class molecular phenotype classification from Whole Slide Images (WSIs) [96, 114], 5-class polyp classification from colonoscopy images [145], cardiac resynchronization therapy response prediction from cardiac MRIs [109], and super resolution of brain MRIs [147]. The remaining articles (n=4) did not include explicit information on whether human baselines and established criteria exist for the envisioned application, e.g., magnification level and nuclei area prediction in breast cancer histology images [90], age estimation in brain MRIs [92], AD status in Diffusion Tensor Images (DTIs), and risk of sudden cardiac death prediction in cardiac MRIs [97]. As previously mentioned, tasks that are routinely performed in clinical evidence may have robust human baselines and clinical guidelines to guide transparent ML development. Applications that are beyond the current possibilities, however, require a more nuanced and human-centered approach that should involve the target end users as early as possible to verify that the assumptions that drive transparency are valid.

Chapter 3

Deep Hiearchical Multi-Label Classification Applied to Chest X-Ray Abnormality Taxonomies

Chest X-Rays (CXRs) are a crucial and extraordinarily common diagnostic tool, leading to heavy research for computer-aided diagnosis (CAD) solutions. However, both high classification accuracy *and* meaningful model predictions that respect and incorporate clinical taxonomies are crucial for CAD usability. To this end, we present a deep hierarchical multi-label classification (HMLC) approach for CXR CAD utilizing the clinical taxonomy. Different than other hierarchical systems, we show that first training the network to model conditional probability directly and then refining it with unconditional probabilities is key in boosting performance. In addition, we also formulate a numerically stable cross-entropy loss function for unconditional probabilities that provides concrete performance improvements. Finally, we demonstrate that HMLC can be an effective means to manage missing or incomplete labels. To the best of our knowledge, we are the first to apply HMLC to medical imaging CAD. We extensively evaluate our approach on detecting abnormality labels from the CXR arm of the Prostate Lung Colorectal and Ovarian (PLCO) dataset, which comprises over $198,000$ manually annotated CXRs. When using complete labels, we report a mean Area Under Curve (AUC) of 0.887, the highest yet reported for this dataset. These results are supported by ancillary experiments on the PadChest dataset, where we also report significant improvements, 1.2% and 4.1% in AUC and average precision, respectively over strong “flat” classifiers. Finally, we demonstrate that our HMLC approach can much better handle incompletely labelled data. These performance improvements, combined with the inherent usefulness of taxonomic predictions, indicate that our approach represents a useful step forward for CXR CAD.

3.1 Clinical Background

Chest X-Rays (CXRs) account for a large proportion of ordered image studies, e.g., in the US it accounted for almost half of ordered studies in 2006 [152]. Commensurate with this importance, CXR *computer-aided diagnosis (CAD)* has received considerable research attention, both prior to the popularity of deep learning [153], and afterwards [154–158]. These efforts have met success and typically approach the problem as a standard multi-label classification scenario, which attempts to make a set of individual binary predictions for each disease pattern under consideration. Truly large-scale CXR classification started with the CXR14 dataset and the corresponding model [154], with many subsequents improvements both in modeling and in dataset collection [157–159]. These improvements include incorporating ensembling [160], attention mechanisms [161–163], and localizations [156, 163–166]. A commonality between these prior approaches is that they typically treat each label as an independent prediction, which is commonly referred to as binary relevance (BR) learning within the multi-label classification field [167]. However, prior work has well articulated the limitations of BR learning [168]. A notable exception to this trend is [155], which modeled correlations between labels using a recurrent neural network. Yet, pushing raw performance further will likely require models that depart from standard multi-label classifiers. For instance, despite their importance to clinical understanding and interpretation [169–171], taxonomies of disease patterns are not typically incorporated into CXR CAD systems, or for other medical CAD domains for that matter. This observation motivates our work, which uses *hierarchical multi-label classification (HMLC)* to both push raw Area Under Curve (AUC) performance further and also to provide more meaningful predictions that leverage clinical taxonomies.

Organizing diagnoses or observations into ontologies and/or taxonomies is crucial within radiology, e.g., RadLex [172], with CXR interpretation being no exception [173–175]. This importance should also be reflected within CAD systems. For instance, when uncertain about fine-level predictions, e.g., *nodules* vs. *masses*, a CAD system should still be able to provide meaningful parent-level predictions, e.g., *pulmonary nodules and masses*. This parent prediction may be all the clinician is interested in anyway. Another important benefit is that observations are conditioned upon their parent being true, allowing fine-level predictors to focus solely on discriminating between siblings rather than on having to discriminate across all possible conditions. This can help improve classification performance [176]. In addition, incorporating taxonomy through hierarchical classification has been well-studied for natural image classification. Prior to the emergence of deep learning, seminal approaches used hierarchical and multi-label generalizations of classic algorithms [177–180]. With the advent of deep learning, a more recent focus has been on adapting deep networks, typically Convolutional Neural Networks (CNNs), for hierarchical classification [181–185]. Interestingly, [178] use an approach similar to popular approaches seen in more recent deep hierarchical *multi-class* classification of natural images [181–183], i.e., train classifiers to predict conditional probabilities at each node. Within medical imaging, there is work on HMLC medical image retrieval using either nearest-neighbor or multi-layer perceptrons [186] or decision trees [175]. However, hierarchical classifiers have not received much attention for medical imaging *CAD* and deep HMLC approaches have not been explored at all.

Elegantly addressing the problem of incompletely labelled data is another benefit of incorporating taxonomy. To see this, note that many CXR datasets are collected using Natural Language Processing (NLP) approaches applied to hospital picture archiving and communication systems (PACSs) [154, 157]. This is a trend that will surely increase given that PACSs remain the most viable source of large-scale medical data [187, 188]. In such cases, it may not always be possible to extract fine-grained labels with confidence. For instance, imaging conditions may have only allowed a radiologist to report "opacity", instead of a more specific observation of "infiltration" vs. "atelectasis". Added to this inherent uncertainty is the fact that NLP approaches for CXR label extraction themselves can suffer from considerable levels of error and uncertainty [157, 189]. As a result, it is likely that CAD systems will increasingly be faced with incompletely labelled data, where data instances may be missing fine-grained labels, but still retain labels higher up in the clinical taxonomy. An HMLC approach can naturally handle such incompletely labelled data. Within the computer vision and text mining literature, there is a rich body of work on handling partial labels [190–197]. When missing labels are positive examples, this problem has also been called positive and unlabelled (PU) learning. Seminal PU works focus on multi-class learning [193, 194]. There are also efforts for *multi-label* PU learning [190–192, 195–197], which attempt to exploit label dependencies and correlations to overcome missing annotations. However, many of these approaches do not scale well with large-scale data [191]. [190] and [191] provide two exceptions to this, tackling large-scale numbers of labels and data instances, respectively. In our case, we are only interested in the latter, as the number of observable CXR disease patterns remains manageable.

For these reasons, we present a deep HMLC approach for CXR CAD. We extensively evaluate our HMLC approach on the CXR arm of the *Prostate Lung Colorectal and Ovarian (PLCO)* dataset [198] with supporting experiments on the PadChest dataset [158]. Experiments demonstrate that our HMLC approach can push raw performance higher compared to both leading “flat” classification baselines and other HMLC alternatives. We also demonstrate that our HMLC approach can robustly handle extremely large proportions of incompletely labelled data with much less performance loss than alternatives. To the best of our knowledge, we are the first to outline an HMLC CAD system for medical imaging and the first to characterize performance when faced with incompletely labelled data.

3.1.1 Contributions

Based on the above, the contributions of our work can be summarized as follows:

- Like other deep hierarchical *multi-class* classifiers, we train a classifier to predict conditional probabilities. However, we operate in the *multi-label* space and we also demonstrate that a second fine-tuning stage, trained using unconditional probabilities, can boost performance for CXR classification even further.

- To handle the unstable multiplication of prediction outputs seen in unconditional probabilities we introduce and formulate a numerically stable and principled loss function.

- Using our two-stage approach, we are the first to apply hierarchical multi-label classification (HMLC) to CXR CAD. Our straightforward, but effective, HMLC approach results in the highest mean AUC value yet reported for the PLCO dataset.

- In addition, we demonstrate how HMLC can serve as an effective means to handle incompletely labelled data. We are the first to characterize CXR classification performance under this scenario, and experiments demonstrate how HMLC can garner even greater boosts in classification performance.

3.2 Materials and Methods

We introduce a two-stage method for CXR HMLC. We first outline the datasets and taxonomy we use in Section 3.2.1 and then overview the general concept of HMLC in Section 3.2.2. This is followed by Sections 3.2.3 and 3.2.4, which detail our two training stages that use conditional probability and a numerically stable unconditional probability formulation, respectively.

3.2.1 Datasets and Taxonomy

The first step in creating an HMLC system is to create the label taxonomy. In this work, our main results focus on the labels and data found within the CXR arm of the PLCO dataset [198], a large-scale lung cancer screening trial that collected 198 000 CXRs with image-based annotations of abnormalities obtained from multiple US clinical centers. While other large-scale datasets [154, 157–159] are *extraordinarily valuable*, their labels are generated by using NLP to extract mentioned disease patterns from radiological reports found in hospital PACSs. While medical NLP has made great strides in recent years, it still remains an active field of research, e.g., NegBio still reports limitations with uncertainty detection, double-negation, and missed positive findings for certain CXR terms [199]. However, irrespective of the NLP's level of accuracy, there are more inherent limitations to using text-mined labels. Namely, examining a text report is no substitute for visually examining the actual radiological scan, as the text of an individual report is not a complete description of the CXR study in question. Thus, terms may not be mentioned, e.g., "no change", even though they are indeed visually apparent. Additionally, a radiologist will consider lab tests, prior radiological studies, and the patient's records when writing up a report. Thus, mentioned terms, and their meaning, may well be influenced by factors that are not visually apparent. Compounding this, text which is unambiguous given the patient's records and radiological studies may be highly ambiguous when only considering text

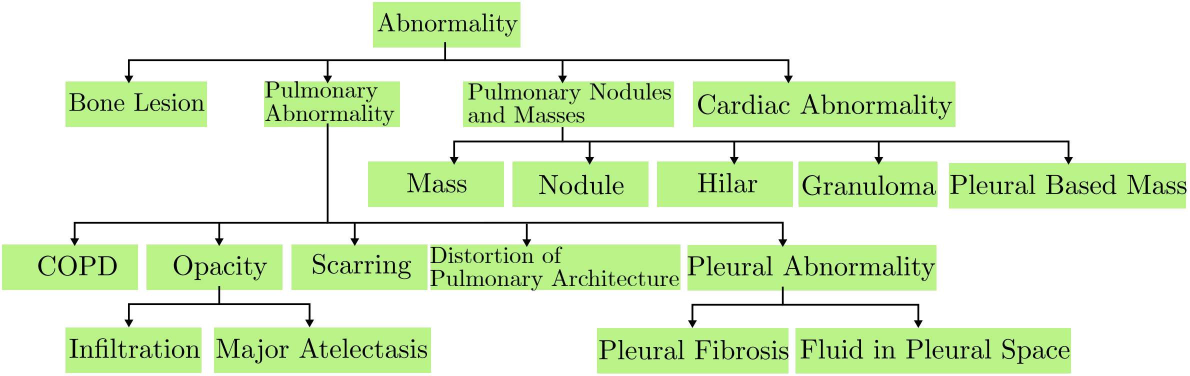


Figure 3-1. Constructed label hierarchy from the PLCO dataset.

alone, e.g., whether a pneumothorax is untreated or not [200]. Indeed, the authors of the PadChest dataset bring up some of these caveats themselves, which are relevant even for the 27% of their radiological reports that are text-mined by hand, which presumably have no NLP errors [158]. An independent study of CXR14 [154] concludes that its labels have low positive predictive value and argues that visual inspection is necessary to create radiological datasets [200]. Consequently, PLCO is unique in that it is the only large-scale CXR dataset with labels generated via *visual observation* from radiologists. Although the PLCO data is older than alternatives [154, 157–159], it has greater label reliability.

Radiologists in the PLCO trial labelled 15 disease patterns, which we call "leaf labels" in our taxonomy. Because of low prevalance, we merged "left hilar abnormality" and "right hilar abnormality" into "hilar abnormality", resulting in 14 labels. From the leaf nodes, we constructed the label taxonomy shown in Figure 3-1. The hierarchical structure follows the PLCO trial's division of "suspicious for cancer" disease patterns vs. not, and is further partitioned using common groupings [173], totalling 19 leaf and non-leaf labels. While care was taken in constructing the taxonomy and we aimed for clinical usefulness, we make no specific claim as such. We instead use the taxonomy to explore the benefits of HMLC, stressing that our approach is general enough to incorporate any appropriate taxonomy. Figure 3-2 visually depicts examples from our

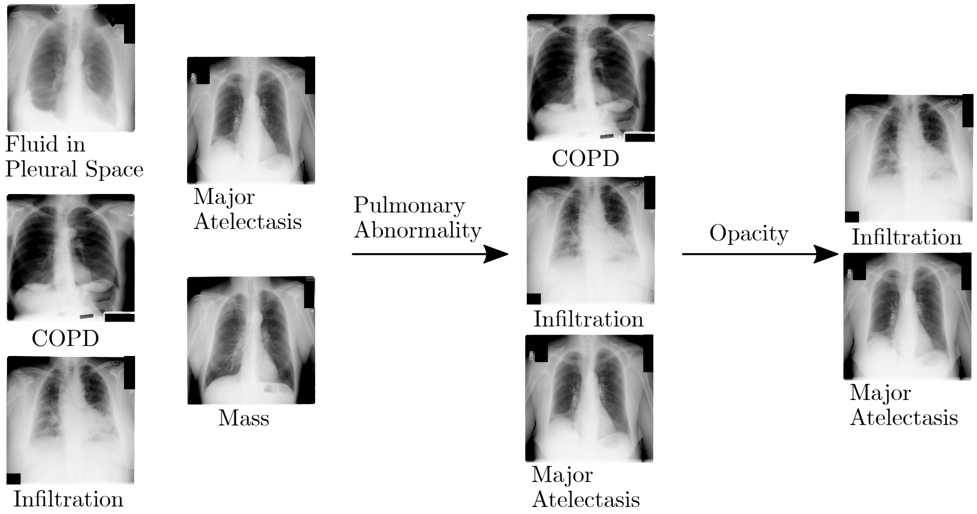


Figure 3-2. Example PLCO CXRs drawn from three levels of our taxonomy. On the left, at the higest level of taxonomy, i.e., "Abnormality", disease patterns may manifest as a variety of visual features within the lung parenchyma, lung pleura, or the surrounding organs/tissues. As one progresses down the taxonomy, i.e., to "Opacity", the discriminating task is narrowed into identifying the "cloudy" patterns seen in both "Infiltration" and "Major Atelectasis."

chosen CXR taxonomy.

As supporting validation to our main PLCO experiments, we also validate on the PadChest dataset [158], which contains 160, 845 CXRs whose labels are drawn from either manual or automatic extraction from radiological *text reports.* We focus on labels categorized as "radiological findings", which are more likely to correspond to actual disease patterns found on the CXRs [158]. Any CXR with a solitary "Unchanged" label is removed, resulting in 121, 242 samples. Uniquely, PadChest offers a complete hierarchical structure for all labels. We remove labels with less than 100 manually labelled samples and only retain labels that align with our PLCO taxonomy. This both ensures we have enough statistical power for evaluation and that we are retaining PLCO-like terms that we can confidently treat as clinically significant. As a result, total 30 out of 191 labels are selected, and our supplementary includes more details of the included and excluded labels. The resulting taxonomy is shown in Figure 3-3. Unlike PLCO, certain parent labels can be positive with no positive children labels,

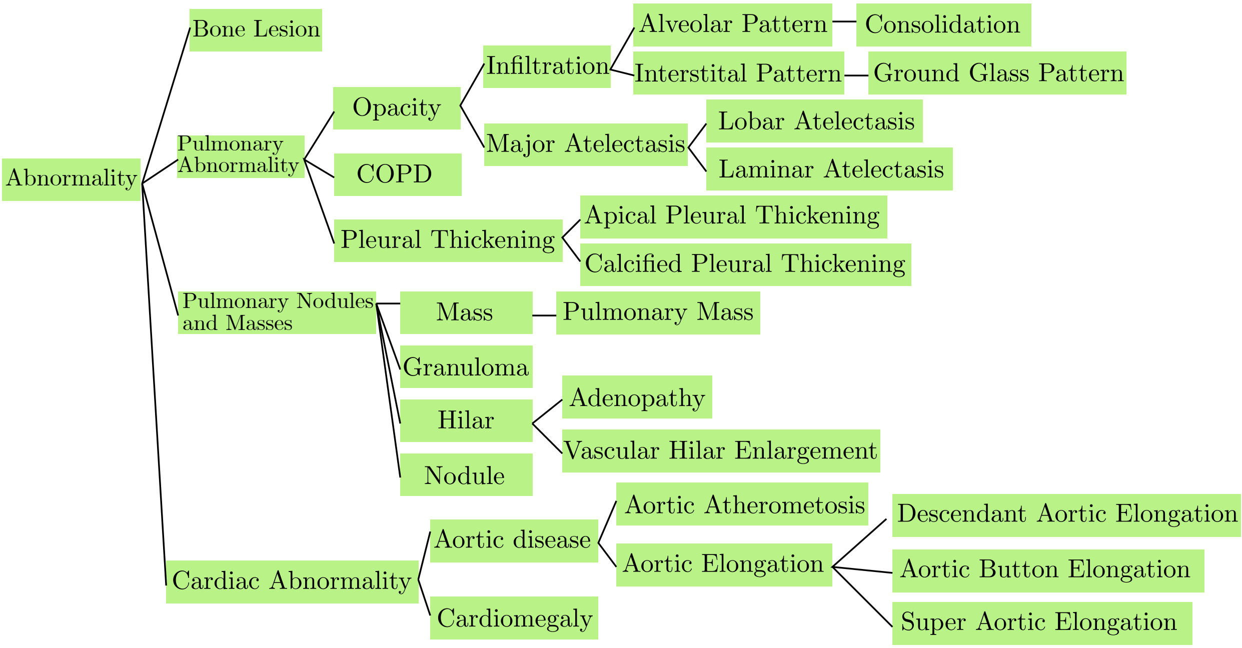


Figure 3-3. Constructed label hierarchy from the PadChest dataset.

e.g., “Aortic Elongation”.

3.2.2 Hierarchical Multi-Label Classification

With a taxonomy established, a hierarchical approach to classification must be established. Because this is a multi-label setting, all or none of the labels in Figure 3-1 can be positive. The only restriction is that if a child is positive, its parent must be too. Siblings are not mutually exclusive. For PLCO, we assume that each image is associated with a set of ground-truth leaf labels and their antecedents, i.e., there are no incomplete paths. However, for PadChest a ground-truth path may terminate before a leaf node. A training set, may have missing labels.

We use a DenseNet-121 [201] model as a backbone. If we use k to denote the total number of leaf and non-leaf labels, we connect k fully connected layers to the backbone’s last feature layer to extract k scalar outputs. Each output is assumed to represent the conditional probability (or its logit) given its parent is true. Thus, once the model is successfully trained, unconditional probabilities can be calculated from the output using the chain rule, e.g., from the PLCO taxonomy the unconditional probability of *scarring* can be calculated as

$$P(\text{Scar.}) = P(\text{Abn.})P(\text{Pulm.}|\text{Abn.})P(\text{Scar.}|\text{Pulm.}), \tag{3.1}$$

where we use abbreviations for the sake of typesetting. In this way, the predicted unconditional probability of a parent label is guaranteed to be greater than or equal to its children labels. We refer to the conditional probability in a label hierarchy as Hierarchical Label Vonditional Probability (HLCP), and the unconditional probability calculated following the chain rule as Hierarchical Label Unconditional Probability (HLUP). The network outputs can be trained either conditionally or unconditionally, which we outline in the next two sections.

3.2.3 Training with Conditional Probability

Similar to prior work [181–183], in the first stage of the proposed training scheme, each classifier is only trained on data conditioned upon its parent label being positive. Thus, training directly models the conditional probability. The shared part of the classifiers, i.e., feature layers from the backbone network, is trained jointly by all the tasks. Specifically, for each image the losses are only calculated on labels whose parent label is also positive. For example, and once again using the PLCO taxonomy, when an image with positive *Scarring* and no other positive labels is fed into training, only the losses of *Abnormality* and the children labels of *Pulmonary Abnormality* and *Abnormality* are calculated and used for training.

Figure 3-4 (a) illustrates this training regimen, which we denote HLCP training. In this work, we use cross entropy (CE) loss to train the conditional probabilities, which can be written as

$$L_{HLCP} = \sum_{m \in M} CE\left(z_m, \hat{z}_m\right) * 1_{\{z_{a(m)}=1\}}, \tag{3.2}$$

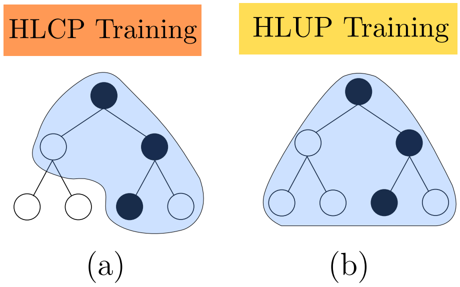


Figure 3-4. The HLCP and HLUP losses are depicted in (a) and (b), respectively, where black and white points are positive and negative labels, respectively. Blue areas indicate the activation area in the loss functions.

where M denotes the set of all disease patterns, and m and $a(m)$ denote a disease pattern and its ancestor, respectively. Here $CE(\cdot, \cdot)$ denotes the cross entropy loss, and $z_m \in \{0, 1\}$ denotes the ground truth label of m, with $\hat{z}_m$ corresponding to the network's sigmoid output.

Training with conditional probability is a very effective initialization step, as it concentrates the modeling power solely on discriminating siblings under the same parent label, rather than having to discriminate across all labels, which eases convergence and reduces confounding factors. It also alleviates the problem of low label prevalence because fewer negative samples are used for each label.

3.2.4 Fine Tuning with Unconditional Probability

In the second stage, we finetune the model using an HLUP CE loss. This stage aims at improving the accuracy of unconditional probability predictions, which is what is actually used during inference and is thus critical to classification performance. Another important advantage is that the final linear layer sees more negative samples. Predicted unconditional probabilities for label m, denoted $\hat{p}_m$, are calculated using the chain rule:

$$\hat{p}_m = \prod_{m' \in A(m)} \hat{z}_{m'}, \tag{3.3}$$

where $A(m)$ is the union of label m and its antecedents. When training using unconditional probabilities, the loss is calculated on every classifier output for every data instance. Thus, the HLUP CE loss for each image is simply

$$L_{HLUP} = \sum_{m \in M} CE\left(z_m, \hat{p}_m\right). \tag{3.4}$$

Figure 3-4(b) visually depicts this loss.

A naive way to calculate (3.4) would be a direct calculation. However, such an approach introduces instability during optimization, as the training would have to minimize the product of network outputs. In addition, the product of probability values within $[0, 1]$ can cause arithmetic underflow. For this reason, we derive a numerically stable formulation below.

Denoting the network's output logits as $\hat{y}_{(.)}$, the predicted unconditional probability of label m can be written as:

$$\hat{p}_m = \prod_{m'} \frac{1}{1 + \exp(-y_{m'})}, \tag{3.5}$$

where we use m' to denote $m' \in A(m)$ for notational simplicity.

The HLUP CE loss is calculated as:

$$\begin{aligned} L_{HLUP} &= - z_m \log(\hat{p}_m) - (1 - z_m)\log(1 - \hat{p}_m), \qquad (3.6)\\ &= - z_m \log\left(\prod_{m'} \frac{1}{1 + \exp(-y_{m'})}\right) \\ &\quad - (1 - z_m)\log\left(1 - \left(\prod_{m'} \frac{1}{1 + \exp(-y_{m'})}\right)\right), \qquad (3.7) \end{aligned}$$

where z_m is the ground truth label of m.

The formulation in (3.7) closely resembles several cross-entropy loss terms combined together. To see this, we can break up the second term in (3.7) to produce the following formulation:

$$\begin{aligned} L_{HLUP} = &- z_m \log\left(\prod_{m'} \frac{1}{1 + \exp(-y_{m'})}\right) \\ &- (1 - z_m)\log\left(\prod_{m'}\left(1 - \frac{1}{1 + \exp(-y_{m'})}\right)\right) + \gamma, \end{aligned} \tag{3.8}$$

where γ is a scalar quantity that must be formulated. The log terms above can then be decomposed as

$$\begin{aligned} L_{HLUP} =& \sum_{m'} \left(-z_m \log \left(\frac{1}{1+\exp(-y_{m'})} \right) \right. \\ & \left. -(1-z_m) \log \left(1 - \frac{1}{1+\exp(-y_{m'})} \right) \right) + \gamma, \quad (3.9) \\ =& \sum_{m'} \ell_{m'} + \gamma, \quad (3.10) \end{aligned}$$

where ℓ_m are individual cross entropy terms, using z_m and $y_{m'}$ as the ground truth and logit input, respectively. Note that (3.10) allows us to take advantage of numerically stable CE implementations to calculate $\sum_{m'} \ell_{m'}$. However to satisfy (3.10), we will need γ to satisfy:

$$\begin{aligned} \gamma =& (1-z_m) \log \left(\prod_{m'} \left(1 - \frac{1}{1+\exp(-y_{m'})} \right) \right) \\ & - (1-z_m) \log \left(1 - \left(\prod_{m'} \frac{1}{1+\exp(-y_{m'})} \right) \right), \quad (3.11) \\ =& (1-z_m) \log \left(\frac{\prod_{m'} \exp(-y_{m'})}{\prod_{m'}(1+\exp(-y_{m'}))} \right) \\ & - (1-z_m) \log \left(\frac{\prod_{m'}(1+\exp(-y_{m'})) - 1}{\prod_{m'}(1+\exp(-y_{m'}))} \right), \quad (3.12) \\ =& (1-z_m) \log \left(\frac{\exp(\sum_{m'} -y_{m'})}{\prod_{m'}(1+\exp(-y_{m'})) - 1} \right), \quad (3.13) \\ =& (1-z_m) \left(\sum_{m'} -y_{m'} - \log \left(\prod_{m'} (1+\exp(-y_{m'})) - 1 \right) \right). \quad (3.14) \end{aligned}$$

If the product within the log-term of (3.14) is expanded, with 1 subtracted, it will result in

$$\gamma = (1-z_m) \left(\sum_{m'} -y_{m'} - \log \left(\sum_{S \in \mathcal{P}(A(m)) \setminus \{\emptyset\}} \exp \left(\sum_{j \in S} -y_j \right) \right) \right), \quad (3.15)$$

where S enumerates all possible subsets of the powerset of $A(m)$, excluding the empty set. For example if there were two logits, y_1 and y_2, the summation inside the log would be:

$$\exp(-y_1) + \exp(-y_2) + \exp(-y_1 - y_2). \quad (3.16)$$

The expression in (3.15) can be written as

$$\gamma = (1 - z_m)\left(\sum_{m'} -y_{m'} - LSE\left(\left\{\sum_{j\in S} -y_j \quad \forall S \in \mathcal{P}(A(m)) \setminus \{\emptyset\}\right\}\right)\right), \tag{3.17}$$

where LSE is the LogSumExp function. Numerically stable implementations of the LogSumExp, and its gradient, are well known. By substituting (3.17) into (3.10), a numerically stable version of the HLUP CE loss can be calculated.

Enumerating the powerset produces an obvious combinatorial explosion. However, for smaller-scale hierarchies, like that in Figure 3-1, it remains tractable. For larger hierarchies, an $O(|A(m)|)$ solution involves simply interpreting the LogSumExp as a smooth approximation to the maximum function, which we provide here for completeness:

$$\gamma \approx (1 - z_m)\left(\sum_{m'} -y_{m'} - \max\left(\left\{\sum_{j\in S} -y_j \quad \forall S \in \mathcal{P}(A(m)) \setminus \{\emptyset\}\right\}\right)\right), \tag{3.18}$$

$$= \begin{cases} (1 - z_m)\left(\sum_{m'} -y_{m'} - \sum_{j:y_j<0} -y_j\right), & \text{if } \exists\, y_{m'} < 0 \\ (1 - z_m)\left(\sum_{m'} -y_{m'} - \max(\{-y_{m'}\})\right), & \text{otherwise} \end{cases}. \tag{3.19}$$

3.3 Experimental

We perform two types of experiments to validate our HMLC approach. The first uses the standard completely labelled setup, helping to reveal how our use of taxonomic classification can help produce better raw classification performance than typical "flat" classifiers. The second uses incompletely labelled data under controlled scenarios to show how our HMLC approach can naturally handle such data, achieving even higher boosts in relative performance.

3.3.1 Complete Labels

Experimental Setup We test our HMLC approach on both the PLCO [198] and PadChest [158] datasets, using the taxonomies of Figure 3-1 and Figure 3-3, respectively. Our emphasis is on PLCO due to its more reliable labels, but evaluations on PadChest provide important experimental support, especially given its larger taxonomy. Following accepted practices in large-scale CXR classification [154, 157, 158], we split the data into single training, validation, and test sets, corresponding to 70%, 10%, and 20% of the data, respectively. Data is split at the patient level, and care was taken to balance the prevalence of each disease pattern as much as possible. As mentioned above, our HMLC approach uses a trunk network, with a final fully-connected layer outputting logit values for each of the nodes of our chosen taxonomy. Our chosen network is DenseNet-121 [201], implemented using TensorFlow. We first train with the HLCP CE loss of (3.2) fine-tuning from a model pretrained from ImageNet [202]. We refer to this model simply as *HLCP*. To produce our final model, we then finetune the HLCP model using the HLUP CE loss of (3.4). We denote this final model as *HLUP-finetune*.

Comparisons In addition to comparing against HLCP, we also compare against three other baseline models, all using the same trunk network fine-tuned from ImageNet pretrained weights. The first, denoted *BR-leaf*, is trained using CE loss on the 14 fine-grained labels. This measures performance using a standard multi-label BR approach. The second, denoted *BR-all* is very similar, but trains a CE loss on all labels independently, including non-leaf ones. In this way, *BR-all* measures performance when one wishes to naively output non-leaf abnormality nodes, without considering label taxonomy. Finally, we also test against a model trained using the HLUP CE loss directly from ImageNet weights, rather than finetuning from the HLCP model. As such, this baseline, denoted *HLUP*, helps reveal the impact of using a two-stage approach vs. simply training an HLUP classifier in one step. For all tested models, extensive hyper-parameter searches were performed on the NVIDIA cluster to optimize mean validation fine-grained AUCs.

For comparisons to external models, we also compare to a recent DenseNet121 BR approach [156] trained on the PLCO data. But, we stress that direct comparisons of numbers are impossible, as [156] used different data splits and only evaluated on 12 fine-grained labels. In the interest of fairness we compare against both (a) their best reported numbers when only training a classifier on CXR disease patterns and (b) their best reported numbers overall, in which the authors incorporated segmentation and localization cues. For (a), we use numbers reported on an earlier work [203], which were higher. Unfortunately, both sets of their reported numbers are based on training data that also included the ChestXRay14 dataset [162], providing an additional confounding factor that hampers any direct comparison.

Finally, we also run experiments to compare our numerically stable implementation of HLUP CE loss in (3.8) to: (a) the naive approach of directly optimizing (3.3); and (b) to a recent rescaling approximation, originally introduced for the multiplication of independent, rather than conditional probabilities, seen in multi-instance learning [165]. This latter approach re-scales each individual probability multiplicand (term) in (3.3) to guarantee that the product is greater than or equal to 1e-7. Similar to the naive approach, the product is then optimized directly using CE loss. For the PLCO dataset, based on a maximum depth of four for the taxonomy, we implement this approach by re-scaling each multiplicand in (3.3) to $[0.02, 1]$.

Evaluation Metrics We evaluate our approach using AUC and average precision (AP), calculated across both leaf and non-leaf labels, when applicable. Additionally, we also evaluate using conditional AUC and AP metrics, which are metrics that reflect the complicated evaluation space of multi-label classification. In short, because more than one label can be positive, multi-label classification performance has exponentially more facets for evaluation than single-label or even multi-class settings. Conditional metrics are one such facet, that focus on model performance conditioned on certain non-leaf labels being positive. Here, we restrict our focus to CXRs exhibiting one or more disease patterns, i.e., *abnormality* being positive. As such, this sheds light on model performance when it may be critical to discriminate what combination of disease patterns are present, which is crucial for proper CXR interpretation [173].

3.3.2 Incomplete Labels

Experimental Setup We also use the PLCO dataset [198] to characterize the benefits of our HMLC approach when faced with incomplete labels. However, after publication of our original work [24], the PLCO organizers altered their data release policies and only released a subset of the original dataset, containing 88 737 labeled CXRs from 24 997 patients. For this reason, we perform our incomplete labels experiments on this smaller dataset, splitting and preparing the data in an identical manner as described in Section 3.3.1.

To simulate a scenario where learning algorithms may be faced with incomplete labels, we removed known labels from the training set using the following controlled scheme:

1. We choose a base deletion probability, $\beta \in [0, 1]$.

2. For data instances with positive labels for “Pleural Abnormality”, “Opacity”, and “Pulmonary Nodules and Masses”, we delete all their children labels with a probability of β. For example, if we delete the children labels of a positive “Pleural Abnormality” instance, then it is no longer known whether the “Pleural Abnormality” label corresponds to “Pleural Fibrosis”, or “Fluid in Pleural Space”, or both.

3. We perform the same steps for data instances with positive labels for “Pulmonary Abnormality” and “Abnormality”, except with probabilities of 0.3β and $0.3^2\beta$, respectively. For example, if the children of a positive instance of “Abnormality” were deleted, then it is only known there are one or more disease patterns present, but not which one(s).

4. A higher-level deletion overrides any decision(s) at finer levels.

5. Because of their extremely low prevalence, we ignore the "Major Atelectasis" and "Distortion in Pulmonary Architecture" labels in training and evaluation.

Note that this scheme makes it more likely to have a missing fine-grained label over a higher-level label, which we posit follows most scenarios producing incomplete labels. When labels are deleted, we treat them as unknown and do not execute any training loss on them. We test our HMLC algorithm and baselines on the following β values: $\{0, .1, .2, .3, .4, .5, .6, .7\}$, which ranges from no incompleteness to roughly 70% of fine-grained labels being deleted. To allow for stable comparisons across β values, we also ensure that if a label was deleted at a certain value of β, it will also be deleted at all higher values of β. To ease reproducibility, we publicly release our data splits (https://github.com/hchen135/Hierarchical-Multi-Label-Classification-X-Rays). All other implementation details are also identical to that of Section 3.3.1.

Evaluation Metrics and Comparisons We measure AUC values and compare our chosen model of HLUP finetune against BR-leaf and BR-all.

3.4 Results and Discussion

We focus in turn on experiments with complete and incomplete labels, which can be found in Section 3.4.1 and Section 3.4.2, respectively.

3.4.1 Complete Labels

Our complete labels experiments first focus on the benefits of our HLUP-finetune approach compared to alternative "flat" and HMLC strategies. Then, we discuss results specifically focusing on our numerically stable HLUP CE loss.

Table 3-I. *PLCO* AUC and AP values across tested models. Mean values across leaf and non-leaf disease patterns are shown, as well as for leaf labels conditioned on one or more abnormalities being present.

	Leaf labels		**Non-leaf labels**		**Leaf labels conditioned on abnormality**	
	AUC	AP	AUC	AP	AUC	AP
[203]	0.865	N/A	N/A	N/A	N/A	N/A
[156]	0.883	N/A	N/A	N/A	N/A	N/A
BR-leaf	0.871	0.234	N/A	N/A	0.806	0.334
BR-all	0.867	0.221	0.852	0.440	0.808	0.323
HLUP	0.872	0.214	0.856	0.436	0.799	0.288
HLCP	0.879	0.229	0.857	0.440	0.822	0.329
HLUP-finetune	**0.887**	**0.250**	**0.866**	**0.460**	**0.832**	**0.342**

3.4.1.1 HLUP-finetune Performance

Table 3-I outlines the PLCO results of our HLUP-finetune approach vs. competitors. As the table demonstrates, the standard baseline BR-leaf model produces high AUC scores, in line with prior work [203]; however, it does not provide high-level predictions based on a taxonomy. Naively executing BR training on the entire taxonomy, i.e., the BR-all model, does not improve performance. This indicates that if not properly incorporated, the label taxonomy does not benefit performance.

In contrast, the HLCP model is indeed able to match BR-leaf's performance on the fine-grained labels, despite also being able to provide high-level predictions. HLUP-finetune goes further by exceeding BR-leaf's fine-grained performance, demonstrating that our two-stage training process can produce tangible improvements. This is underscored when comparing HLUP-finetune with HLUP, which highlights that without the two-stage training, HLUP training cannot reach the same performance. If we limit ourselves to models incorporating the entire taxonomy, our final HLUP-finetune model outperforms BR-all by 2% and 2.9% in leaf-label mean AUC and AP values, respectively. Because HLUP-finetune shares the same labels as BR-all, the performance boosts of the former over the latter demonstrate that the additional

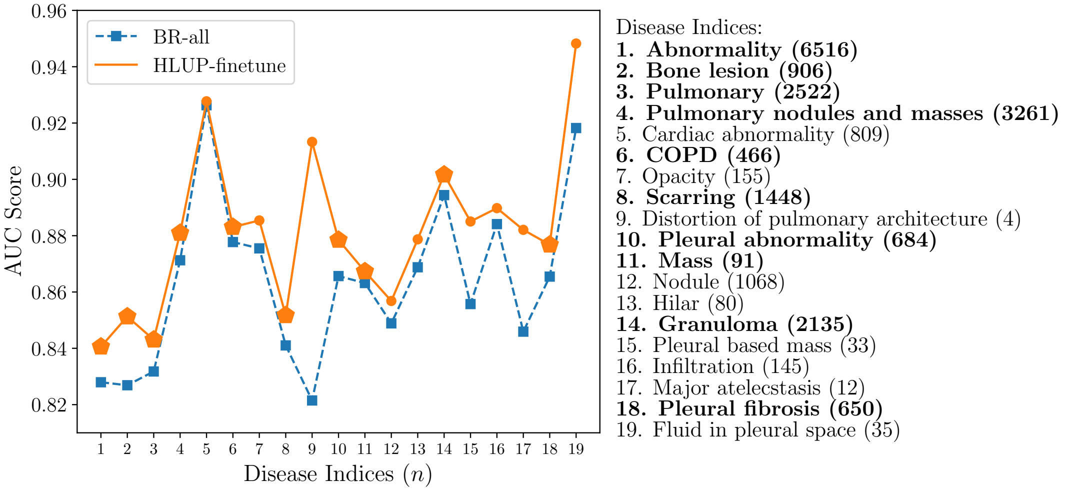


Figure 3-5. Comparison of AUC scores for all fine-grained and high-level (non-leaf) disease patterns for the BR-all and HLUP-finetune models. The dashed line separates the fine-grained from the high-level (non-leaf) disease patterns. Boldface labels and larger graph markers denote disease patterns exhibiting statistically significant improvement ($p < 0.05$) using the StAR software implementation [1] of the non-parametric test of [2].

output nodes seen in HMLC are not responsible for performance increases. Instead, it is indeed the explicit incorporation of taxonomic structure that leads to improved performance.

Figure 3-5 provides more details on these improvements, demonstrating that AUC values are higher for HLUP-finetune compared to the baseline method for all fine-grained and high-level disease patterns. Interested readers can find these AUC values in our supplementary materials. Although not graphed here for clarity reasons, HLUP-finetune also outperformed the HLCP method for all disease patterns. Of note is that statistically significant differences also respect the disease hierarchy, and if a child disease pattern demonstrates statistically significant improvement, so does its parent.

Of particular note, when considering AUCs conditioned on one or more abnormalities being present (last column of Table 3-I), the gap between all HMLC approaches and "flat" classifiers increases even more. As can be seen in such settings, HLUP-finetune

Table 3-II. *PadChest* AUC and AP values across tested models. Mean values across leaf and non-leaf disease patterns are shown, as well as for leaf labels conditioned on one or more abnormalities being present.

	Leaf labels		**Non-leaf labels**		**Leaf labels conditioned on abnormality**	
	AUC	AP	AUC	AP	AUC	AP
BR-leaf	0.825	0.104	N/A	N/A	0.743	0.212
BR-all	0.825	0.110	0.820	0.221	0.739	0.204
HLUP	0.831	0.114	0.828	0.220	0.752	0.211
HLCP	0.831	0.135	0.833	0.240	0.765	0.244
HLUP-finetune	**0.837**	**0.145**	**0.840**	**0.253**	**0.778**	**0.261**

still exhibits increased performance over the baseline models and also the next-best hierarchical model. Importantly, if we compare the conditional AUCs between BR-all and HLUP-finetune, we see a 2.4% increase. This indicates that HMLC is particularly effective at differentiating the exact combination of abnormalities present within an image. This may reduce the amount of spurious and distracting predictions upon deployment, but more investigation is required to quantify this.

We also note that HLUP-finetune managed to outperform [156]'s AUC numbers, despite the latter incorporating almost twice the amount of data and also including additional localization and segmentation tasks. However, we again note that [156] used a different data split and only 12 fine-grained labels, so such comparisons can only be taken so far.

Experiments on PadChest further support these results, with trends mirroring that of the PLCO experiments. As can be seen in Table 3-II, HLUP-finetune outperforms both the BR baselines and HMLC alternatives. Moreover, just like the PLCO experiments, when evaluating AUC and AP conditioned on one or more abnormalities being present, the performance gaps between HLUP-finetune and alternatives further increase. The relative performance improvements demonstrate that our HMLC approach generalizes well to a different CXR dataset outside of PLCO, even though

PadChest uses a different taxonomy and was collected with very different patient populations at a much later date.

The PLCO and PadChest performance boosts are in line with prior work that reported improved classification performance when exploiting taxonomy, e.g., for text classification [177, 204], but here we use HMLC in a more modern deep-learning setting and for an imaging-based CAD application. In particular, given that taxonomy and ontology are crucial within medicine, the use of hierarchy is natural. Because the algorithmic approach we take remains very simple, our HMLC approach may be an effective method for many other medical classification tasks outside of CXRs.

The discussion of the performance boosts garnered by HMLC are very important, but it should also be noted that HMLC provides inherent benefits outside of raw classification performance. By ensuring that clinical taxonomy is respected, i.e., a parent label's pseudo-probability will always be greater than or equal to any of its children's, HMLC provides a more interpretable and understandable set of predictions that better match the top-down structure of medical ontology.

In addition to exploring the benefits of the conceptual approach of HMLC to CXR classification, our work also demonstrates that a two-stage HLUP finetuning approach can provide performance boosts over the more common one-stage HLCP training seen in many prior deep-learning works [181–183]. As such, our two-stage approach may also prove useful to hierarchical classifiers seen in other domains, such as computer vision or text classification.

3.4.1.2 Numerically Stable HLUP

Table 3-III demonstrates that our numerically stable HLUP CE loss results in much better AUCs compared to the competitor rescaling approach [165] and to naive HLUP training when starting from ImageNet weights. However, there were no performance improvements when compared to the naive approach when finetuning from the HLCP

Table 3-III. Comparison of AUCs produced using different HLUP CE loss implementations for PLCO.

HLUP (naive)	HLUP (rescale)	HLUP (ours)	HLUP-finetune (naive)	HLUP-finetune (rescale)	HLUP-finetune (ours)
0.864	0.853	0.872	0.886	0.867	0.887

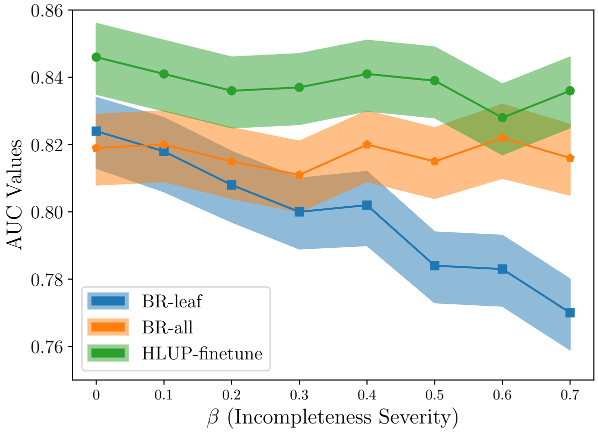


Figure 3-6. Mean AUC scores under different levels of label incompleteness with confidence intervals representing the 2.5th and 97.5th percentiles of 5000 resampling with replacement bootstrap rounds [3].

weights. We hypothesize that the predictions for the HLCP are already at a sufficient quality that the numerical instabilities of the naive HLUP CE loss are not severe enough to impair performance. Nonetheless, given the improvements when training from ImageNet weights, these results indicate that our HLCP CE loss does indeed provide tangible improvements in convergence stability. We expect these improvements to be greater given taxonomies of greater depth, and our formulation should also prove valuable to multi-instance setups which must optimize CE loss over the product of large numbers of probabilities, e.g., the 256 multiplicands seen in [165].

3.4.2 Incomplete Labels

Figure 3-6 shows the results of our incompletely labelled experiments. As can be seen when all labels are present, i.e., $\beta = 0$, the results mirror that of Section 3.4.1, with HLUP-finetune outperforming the baseline models and the BR-all providing no improvements over BR-leaf. As the incompleteness severity increases, BR-leaf's performance drastically drops, while BR-all and HLUP-finetune are much better able to manage label incompleteness. At the highest β level, the performance gap between HLUP-finetune and BR-leaf almost reaches 7%. Per-abnormality AUC values can be found in our supplementary materials.

Our results demonstrate that incorporating hierarchy can be an effective means to manage incomplete labels. Specifically, while HLUP-finetune's performance does indeed drop as the incompleteness severity increases, it does so at a drastically reduced rate compared to the standard BR-leaf classifier. Interestingly, BR-all, which trains all outputs but without incorporating a taxonomy, also manages to retain an equally graceful performance drop. However, HLUP-finetune's roughly 2% AUC performance advantage over BR-all indicates that properly incorporating the taxonomic hierarchy is necessary to boost classification performance. We suspect the anomaly at $\beta = 0.6$ is due to variability caused by the randomness of the training procedure and we reran our experiments at this β value which confirmed this. Ideally, running multiple training runs at each β value would allow us to produce confidence bars that take into account effects from random weight initialization and sampling, but time and computational resources did not allow us to perform this extremely demanding set of experiments. Finally, HLUP-finetune has the added important benefit of producing predictions that respect the taxonomy, which is something that BR-all does not do. Thus, these results indicate that when possible, incorporating a HMLC approach can be an effective means to manage incompletely labelled data. As the prevalence of text-mined PACS medical imaging data increases, we expect the need for approaches to gracefully handle missing labels to increase, and our HMLC approach may provide a useful cornerstore of future work in this direction.

3.5 Conclusions

We have presented a two-stage approach for deep HMLC of CXRs that combines conditional training with an unconditional probability fine-tuning step. To effect the latter, we introduce a new and numerically stable formulation for HLUP CE loss, which we expect would also prove valuable in other training scenarios involving the multiplication of probability predictions, e.g., multi-instance learning. Through comprehensive evaluations, we report the highest mean AUC on the PLCO dataset yet, outperforming hierarchical and non-hierarchical alternatives. Supporting experiments on the PadChest dataset confirm these results. We also show performance improvements conditioned on one or more abnormalities being present, i.e., predicting the specific combination of disease patterns, which is crucial for CXR interpretation. Experiments with incompletely labelled data also demonstrate that our two-stage HMLC approach is an effective means to handle missing labels within training data.

There are several interesting avenues of future work. For instance, while the straightforward HMLC approach we take enjoys the virtue of being easy to implement and tune, it is possible that more sophisticated approaches, e.g., using hierarchical features or dedicated classifiers, may garner even further improvements. Prior work using classic, non deep-learning approaches, explored these options [177–180, 204], and their insights should be applied today. Another important topic of future work should be on incorporating uncertainty within HMLC. This would allow a model, when appropriate, to predict high confidence for non-leaf label predictions but lower confidence for leaf label predictions, enhancing its usefulness in deployment scenarios. Future work should also consider applications outside of CXRs both within and without medical imaging, e.g., genomics or proteomics. Finally, one issue for further investigation is to better understand the implications of the annotation noise described by [156], both for training and for evaluation. Relevant to this work, assessing label noise at higher levels of hierarchy should be an important focus going forward.

Chapter 4

Toward Automated Interpretable AAST Grading for Blunt Splenic Injury

The American Association for the Surgery of Trauma (AAST) splenic organ injury scale (OIS) is the most frequently used CT-based grading system for blunt splenic trauma. However, reported inter-rater agreement is modest, and an algorithm that objectively automates grading based on transparent and verifiable criteria could serve as a high-trust diagnostic aid. To pilot development of an automated interpretable multi-stage deep learning-based system to predict AAST grade from admission trauma CT. Our pipeline includes 4 parts: 1) automated splenic localization, 2) Faster RCNN-based detection of pseudoaneurysms (PSA) and active bleeds (AB), 3) nnU-Net segmentation and quantification of splenic parenchymal disruption (SPD), and 4) a directed graph that infers AAST grades from detection and segmentation results. Training and validation is performed on a dataset of adult patients (age $\geq$ 18) with voxelwise labeling, consensus AAST grading, and hemorrhage related outcome data (n = 174). AAST classification agreement (weighted κ) between automated and consensus AAST grades was substantial (0.79). High grade (IV and V) injuries were predicted with accuracy, PPV, and NPV of 92%, 95%, and 89%. AUC for predicting hemorrhage control intervention was comparable between expert consensus and automated AAST grading (0.83 vs. 0.88). The mean combined inference time for the pipeline was 96.9 seconds. The results of our method were rapid and verifiable, with the high agreement between automated and expert consensus grades. Diagnosis of high-grade lesions and prediction of hemorrhage control intervention produced accurate results in adult patients.

4.1 Clinical Background

Splenic injury is the most common solid organ injury in adult blunt abdominal trauma [26, 27]. It is routinely evaluated on admission CT when abdominal trauma is suspected, both in stable patients, and in those with hemodynamic compromise that demonstrate transient response to fluid resuscitation [205]. In 2018, the American Association for the Surgery of Trauma (AAST) Patient Assessment Committee (PAC) introduced an updated AAST splenic organ injury scale (OIS) for treatment decision making based on admission two-phase abdominopelvic CT examination. While AAST grading has existed as a research tool for decades, the recent update reflects an attempt to operationalize this grading system for point of care use and standardize management practices. Treatment options vary by injury severity, including routine observation for low grade injuries, and urgent angioembolization or splenectomy to control hemorrhage in high-grade injuries. Patient selection is critical as hemorrhage control interventions require mustering limited resources and staff and are not without the potential for both short and long-term morbidity, including from catheter-related complications and an increased lifetime risk of overwhelming post-splenectomy infection (OPSI) [206, 207]. Clinical decision making should be rapid and based on objective criteria, as the spleen is a highly vascular organ and severe splenic injury can potentially lead to exsanguination, multi-organ system failure, and death [34].

CT-based AAST grading enjoys widespread adoption among trauma surgeons, however, in a survey of AAST member practices, only 45% of respondents reported routine use of the AAST grading for blunt splenic trauma by radiologists at their institutions [28]. Even in an ideal circumstance of ubiquitous radiologist adoption and reporting, classification systems are prone to variability in the perceived grades among readers with varying experience and specialization, and reported agreement for the splenic AAST OIS has been modest under research conditions [208, 209]. In practice, radiologists are subject to shifting circumstances in their clinical environment with respect to study volume, reading room distractions, and fatigue-related performance degradation, such as from circadian rhythm disruptions after multiple consecutive night shifts [29–33]. Furthermore, admission trauma CT interpretation is time consuming. Among expert trauma radiologists, interpretation turnaround times for severely injured patients commonly exceed 20-30 minutes [29].

Automated AAST grading could potentially provide a rapid, objective, and accurate second-reader capability, but there has been limited automation research involving the individual CT features of splenic injury and no work has described automated AAST grading to date [210–214].

Black box methods are prone to spurious causal inference [215] and are unaccountable to end-users as the reasoning used to arrive at a decision cannot be verified. For an intelligent system to be considered responsible, ethical, and trustworthy -a requirement in the high-stakes setting of trauma care- it must at a minimum include a layer of explainability to ensure that the decisions made are justifiable [16, 216]. Since AAST grading is a multi-stage process, the intermediate steps of an automated method should be interpretable to end-users, giving them agency to verify individual model assumptions should they choose to do so. Interpretability involves the provision of packets of symbolic information that are semantically similar to the common-sense causal reasoning that would be used by an expert to arrive at a decision- in this case, at a given splenic injury grade [217, 218].

To this end, we leverage transparent deep learning approaches that are based on clinical grading standards to develop a novel automated multi-stage deep learning (DL) method that predicts the AAST splenic OIS using the most salient features of the grading system, namely active bleeding, pseudoaneurysm, and splenic parenchymal disruption [37].

4.2 Materials and Methods

4.2.1 Datasets

The work was conducted as part of an IRB approved study and utilized two deidentified single institution datasets. The primary clinical dataset is previously described [205] and consists of 174 dual phase trauma CT scans from consecutively selected adult patients (age $\geq$ 18) collected between 2017-2019 and archived at 1.5-3 mm section thickness, with voxelwise labeling of pseudoaneurysm (PSA), active bleed (AB), and splenic parenchymal disruption (SPD). PSAs in this context are vascular injuries contained by splenic parenchyma with densities similar to or slightly higher than the blood pool [26]. Foci of AB extend beyond the splenic parenchyma, and typically increase in size on the portal venous phase. The portal venous phase is optimal for delineation of SPD and AB, whereas PSA is best detected on the arterial phase [219, 220].

All studies in this dataset had accompanying AAST consensus grading by three expert trauma radiologists, and outcome data including whether patients underwent angioembolization or splenectomy. Foci of active bleed and pseudoaneurysm occupy a small fraction of an abdominal CT volume when present. To address AB and PSA class imbalance, the dataset was augmented with labeled dual phase CTs from a second existing blunt splenic injury dataset with 68 consecutively selected patients who underwent splenic hemorrhage control intervention and had AB, PSA or both on CT between 2007-2016 [221] (ABPSA dataset). A third dataset with the subset of

41 labeled splenic normals from the medical segmentation decathlon challenge [222] (SMSD dataset) was employed to develop an initial automated localization step of injured spleens.

4.2.2 Summary of the AAST Splenic Organ Injury Scale and Clinical Relevance

All patients with intraperitoneal AB receive a grade of V in this system, while any patient with PSA but no AB receives a grade of IV [37]. This assignment is irrespective of size and number of splenic vascular lesions. High grade (IV and V) injuries are considered to necessitate angioembolization (AE) for hemorrhage control at a minimum, and surgeons may opt instead for early splenectomy. Rates of failure for attempted splenic salvage for high-grade injuries are historically high, ranging from 20% to over 60% [207, 223–226] but are improved with liberal use of angioembolization [224]. The surgeon's judgement and institution-specific guidelines play an important role in the choice of hemorrhage control intervention [225]. In patients without splenic vascular injury, the AAST splenic OIS grade is primarily determined by the extent of visually estimated SPD [37, 227] using diameter measurement cut-offs established using rules of thumb in the original 1994 AAST classification [227]. Management of grade 3 injuries (with greater than approximately 3 cm estimated SPD depth but less than 25% parenchymal involvement) is highly variable. At many institutions, these injuries are considered low grade and are routinely managed conservatively [225], however some investigators report improved salvage rates with routine angiographic screening followed by AE if a vascular injury is seen on the image intensifier [207, 228], and still others recommend the routine use of AE as a precautionary measure [229]. Variability in practice patterns lies in the potential for missed small or subtle vascular injury on CT due to variable scan timing, and transient vessel thrombosis or spasm [229]. Low grade (grade I and II) injuries have less than approximately 3 cm SPD depth.

Conservative management is considered the standard of care for low grade injuries across institutions [225, 226].

4.2.3 Automated Splenic Injury Grading Pipeline: Overview

The complete pipeline for our proposed automated AAST OIS grade prediction method is shown in 4-1. The pipeline begins with an automated 3D cropping step aimed at a) reducing irrelevant background which could otherwise contribute to false positive results and b) increasing the proportion of positive voxels in the data given small target volumes of AB, PSA, and SPD. 2D Faster R-CNN [230] is then applied to the detection of AB on portal venous phase axial slices and PSA on arterial axial slices, leveraging the optimal phase for detection of each feature [219, 220]. SPD is segmented using nnU-Net [231] and quantified using voxel counting. Vascular injury detections and SPD volumes are then fed into a hierarchical rules-based system to derive the predicted AAST grade.

4.2.4 Step 1: Automated Splenic Localization

A semi-supervised method using the Noisy Student Algorithm [232] was employed to derive whole-organ label masks for injured spleens. The method initially utilizes a 3D U-Net trained on the external medical segmentation decathlon challenge (SMSD) “teacher” splenic normal labels to generate pseudo-labels in the inhouse clinical and ABPSA datasets. The labels and pseudo-labels are used to create improved spleen segmentations in an iterative process. The 3D U-Net predicted segmentations from a given iteration are used as ground truth training cases in the next iteration, resulting in gradual segmentation refinement. We used the MONAI platform [233] to construct the 3D U-Net [234], with a learning rate of 1e-4. All training was performed using an RTX 3090 NVIDIA GPU with 64 GB of RAM. Training involved six iterations with 600 epochs per iteration. Subsequently, the splenic volume is dilated by 30 voxels to include relevant perisplenic soft tissue structures and axial images above and below the dilated volume are excluded. Following this pre-processing step, visual inspection of each CT study indicated that all foci of vascular injury and SPD were contained within the cropped range and there were no failures.

4.2.5 Step 2: Vascular Lesion Detection

Following splenic localization, 2D Faster R-CNN [230], a two-stage object detection network, was used to detect PSA on axial arterial phase images, and AB on axial portal venous phase images. The network first extracts image features and generates region proposals. Second, it fine tunes the box proposal size and location and classifies each proposal. We used ResNeXt-101 with Feature Pyramid Network (FPN) as the backbone given best performance in the COCO object detection task [235, 236]. Training was augmented using the non-overlapping ABPSA dataset, within which each CT study includes at least one focus of pseudoaneurysm or active bleed. Faster R-CNN was trained in five-fold cross-validation, splitting the combined dataset evenly into 5 independent subsets to avoid data leak, and using each fold for validation and the remaining 4 folds for training. Thresholds were selected to achieve the highest possible sensitivity. Faster R-CNN was implemented in the detectron2 platform [237], with the following parameters: learning rate of 0.02, a 10x decay at 15,000 iterations, and a total of 30,000 training iterations.

4.2.6 Step 3: Splenic Parenchymal Disruption Segmentation and Quantification

We applied nnU-Net [231] to segment splenic parenchymal disruption (SPD) due to its state-of-the-art performance across a large variety of segmentation tasks. nnU-Net trains four models in five-fold cross-validation (2D U-Net, low resolution 3D U-Net, high resolution 3D U-Net, and a low- and high-resolution cascaded 3D U-

Net) and determines the best performing model or ensemble of models for inference. Design choices including hyperparameter selection and pre-processing steps are made automatically from specific dataset properties known as the dataset and pipeline fingerprint [231]. Voxel counting is then applied to automated label masks to determine laceration volumes.

4.2.7 Step 4: AAST Grade Determination

The automated detections and segmentation-derived volumes are directly applied to a hierarchical AAST OIS-based system of rules (Figure 4-1). The directed graph starts from the highest-grade decision, proceeding toward the lowest grade in a manner similar to how AAST grading is employed in clinical practice (Table 4-I). First, if AB is detected, the patient receives a grade of V. If no AB is detected, but PSA is detected, the patient receives a grade of IV. If no vascular injury is present, the grade is determined by the extent of splenic parenchymal disruption. In our method, grades are stratified by the automated SPD volume in place of visual estimation of injury depth using a logistic regression-derived cut-off. Since low grade lesions are managed conservatively, we combined grades I and II into a single “low grade” class. A laceration volume of 14 mL optimally discriminated between low grade (I and II) and grade III lesions.

4.3 Results

Descriptive statistics for the clinical dataset are provided in Table 4-II. Weighted Cohen’s κ between automated and consensus grades in the clinical dataset of 174 patients was 0.79.

Using radiologist expert consensus grading as the reference standard, diagnosis of high grade (IV and V) splenic injuries- those that require urgent hemorrhage control intervention for splenic salvage [226, 238]- was achieved with an accuracy, sensitivity, specificity, NPV, and PPV of 92%, 93%, 92%, 95%, and 89%. The proposed method correctly identified almost all high-grade injuries. Only 1 of 69 patients with high grade (IV and V) injuries was under-estimated as a grade III injury. 9 grade III injuries were over-estimated as high grade. All patients classified as grade I and II by radiologists (n = 91) were correctly predicted as low-grade injuries using our method. This indicates that among patients who would normally be managed conservatively, there were no false positive severe injuries. In subanalysis of high-grade injuries, 19 out of 51 consensus grade IV patients are over-estimated by our method as grade V, but only 2 out of 18 grade V patients are underestimated as grade IV. The AUC for predicting a composite outcome of intervention with angioembolization or splenectomy was 0.83 for automated grades, comparable to an AUC of 0.88 for consensus grading.

Performance of detection and segmentation tasks. For AB detection in the clinical and ABPSA datasets, Faster R-CNN achieved an AUC, accuracy, sensitivity, specificity, NPV, and PPV of 0.84, 88%, 82%, 91%, 91%, and 83% (Table 4-III). For PSA, Faster RCNN achieved an AUC, accuracy, sensitivity, specificity, NPV, and PPV of 0.79, 83%, 91%, 78%, 94%, and 71%. Examples of AB and PSA box detections are shown in Figure 4-2.

A review of AB false positive detections revealed that in 13 of 19 patients, the detection network misclassified pseudoaneurysm as active bleeding on the portal venous phase, owing to lingering pseudoaneurysm blush which otherwise characteristically washes out and is inconspicuous on this phase [26, 219]. This could be attributable to imperfect CT timing using a descending thoracic aorta ROI trigger threshold or variability in cardiac output between patients [205]. Nevertheless, box proposals of these lesions provide transparent results which would allow radiologists, interventionalists, or surgeons to reject detections they disagree with.

Patients in the clinical dataset had labeled laceration with a range of volumes between 0.1-255.1 mL (median volume: 1.8 mL). All patients with < 1 mL had low grade injuries. Automated segmentations in patients with >= 1 mL SPD and no vascular injury that would take absolute priority in injury grading had volume Similarity (VS) index of 0.68, and dice similarity coefficient of 0.54 with respect to manual labels, corresponding with high-saliency visual results that conformed to the margins of laceration (Figure 4-3). Pearson's r between manual and automated volumes was 0.89 ('excellent' range).

Mean inference times for our method included 1.5 seconds for automated splenic localization, 4.2 seconds for Faster R-CNN, and 91.2 seconds for ensembled nnU-Net, for a total of 96.9 seconds.

4.4 Discussion

The AAST splenic organ injury scale is often used to guide surgical management decisions. High grade (AAST IV and V) lesions typically require angioembolization or splenectomy for hemorrhage control [205, 224–226, 238]. Low grade (AAST I and II) lesions are routinely managed conservatively [225, 226]. Management of grade III lesions remains variable and institution dependent [207, 228, 229]. AAST grading is limited by modest interobserver agreement, inconsistent reporting, and the long interpretation and reporting times of admission trauma CT [28, 29, 208, 209]. An interpretable automated system could augment objective decision-making as a second-reader diagnostic aid, producing verifiable visual results that could be accepted or rejected by the end-user.

To date we are not aware of previous attempts to automate AAST splenic OIS grading, either with black box or interpretable methods. Few studies report automated methods for detection or segmentation of individual features relevant to the grading of splenic injury. Several works describe whole-spleen segmentation in trauma patients using semi-automated [210] and automated methods, such as with 3D active shape contours and probabilistic atlases [211, 212], however whole-spleen volumes are of unknown clinical import in trauma. Other work describes black-box detection but not quantification of splenic parenchymal disruption using a random forest method and a convolutional neural network with a long short term memory (LSTM) model [213]. One group examined detection of active bleed but not pseudoaneurysm using a hand-crafted feature engineering-based method. Of 30 splenic injury subjects, 4 had active bleeding. The method had a detection accuracy and PPV of only 73% and 33% respectively [214]. In more recent work using deep learning segmentation methods, automated liver parenchymal disruption volumes predicted angiopositivity on subsequent conventional angiography [239]. A variety of robust methods have emerged using DL for non-trivial hemorrhage-related tasks, including hemoperitoneum, extraperitoneal pelvic hematoma, and hemothorax quantitative visualization [240–243]. Additional DL-related work in the spleen has demonstrated the feasibility of splenic vascular injury segmentation, however without an initial detection step or the ability to differentiate between pseudoaneurysm and active bleeding [221]. Quantification of vascular injury burden is presently not included in the AAST OIS framework, and detection of vascular injury is made in a binary fashion on the patient level.

In the present work, we pilot development of an interpretable automated AAST splenic grade prediction pipeline using the dual-phase imaging protocol currently recommended by the AAST Patient Assessment Committee [37]. Dual phase imaging is optimized for delineation of splenic disruption and detection of active bleeding on the portal venous phase, and detection of pseudoaneurysm on the arterial phase [219, 220].

Our pipeline begins with robust splenic localization, followed by detection of pseudoaneurysm and active bleed using Faster R-CNN with a RexNeXt-101 backbone, and nnU-Net segmentation of splenic parenchymal disruption. The detections and splenic volumes are then fed into an intuitive rules-based system guided by major concepts of the AAST OIS combined with consideration of clinical evidence and expert knowledge. Using the Landis and Koch scheme [244], we achieved substantial agreement with expert consensus ground truth AAST grading (weighted κ of 0.79), and a 92% accuracy for predicting high grade (IV and V) lesions. The AUC of automated grades for predicting a composite outcome of angioembolization or splenectomy was comparable to prediction of the same outcome using expert consensus AAST grading. The method is much more rapid in inference compared with reported admission trauma CT interpretation times.

For active bleed detection, Faster R-CNN achieved an accuracy, NPV, and PPV of 88%, 91%, and 83%, and for pseudoaneurysm detection, an accuracy, NPV, and PPV of 83%, 94%, and 71%. nnU-Net results demonstrated reasonably high DSC and volume similarity for a task involving small target volumes, with excellent Pearson correlation between manual and automated volumes (0.89) and high-quality visual results. A threshold of 14 mL optimally distinguished between low grade (I and II) and grade III lesions.

Our study had some limitations. In our clinical dataset, we were not able to determine a threshold distinguishing between grade III and the subset of grade IV patients without detected pseudoaneurysms as there were only two such patients, both with volumes that overlapped with the volume distribution of grade III injuries. A larger multi-institutional dataset will contain more such patients and likely allow determination of additional cut-offs for grade IV and V injuries. The 2018 AAST OIS is not without controversy. Despite the recommendation of dual phase scanning, this protocol isn't widely adopted at present. Some high-volume level I trauma centers perform scanning of the abdomen in the portal venous phase only or use a single-phase split bolus protocol [245–247]. Additional features included in the AAST OIS were selected using heuristics without strong evidence. For example, the estimated size of subcapsular hematoma is included, although our review of the literature yielded few studies supporting this is as a univariate predictor of outcome [248], and we are not aware of studies showing that this feature is independently predictive when accounting for SPD, and vascular injury. Laceration and intraparenchymal hematoma (blood pooled within the interstices of a laceration) are typically indistinguishable and grouped as splenic parenchymal injury or disruption in scientific works and clinical practice [219, 249]. The 2018 AAST discriminates between intra- and extraperitoneal active bleeding even though the extensive literature on vascular lesions does not differentiate between active bleeding confined to or extending beyond the capsule into the peritoneal cavity [219, 220, 250, 251]. Additionally, the AAST OIS currently includes capsular tear for discriminating between grade I and II lesions [37]. The capsule is microscopic and not directly visible on CT. Capsular laceration is only implied by the presence of SPD and this feature may be redundant. By including only those features in our simplified system supported by strong evidence [219, 225, 238, 250, 252], we managed to achieve substantial agreement with human expert grading. The internal and clinical validity of each feature of the 2018 AAST splenic OIS is an active area of investigation by the American Society of Emergency Radiology (ASER) Splenic Trauma Expert Panel [253], and we will consider inclusion of additional features as dictated by its ongoing conclusions and further emergence of scientific evidence. Finally, our method is trained and validated using only adult patients, and future work will ultimately need to include pediatric trauma victims.

Other future avenues of investigation may include collaboration with participants of the ASER panel, which has curated a large, as yet unlabeled, multicenter CT dataset with studies performed using a variety of protocols on a wide range of scanner makes and models. The method could be refined and retrained and tested on a hold-out sample. A simulated deployment study comparing multileader agreement and diagnostic performance with and without our interpretable method as a diagnostic support system is also planned.

In conclusion, in this single center pilot study, we developed a rapid interpretable automated method for grading splenic injury using the most salient features of the AAST splenic OIS. The method achieved high agreement with, and accuracy compared to consensus expert AAST grading in cross-validation. Prediction of hemorrhage control intervention was comparable between automated and consensus grading. Future avenues include scaling to a larger dataset, conducting a simulated deployment study, and assessing user acceptance.

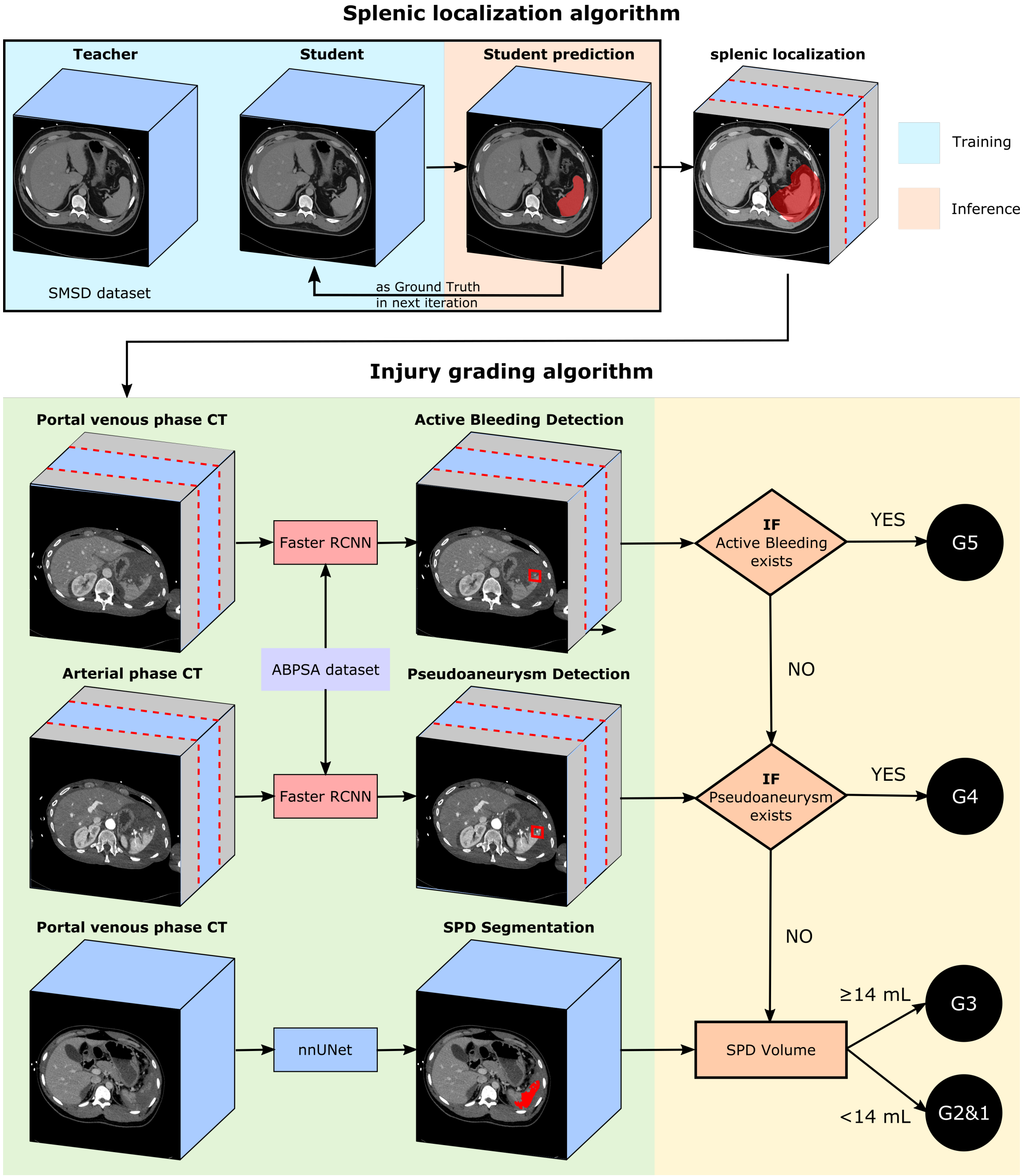


Figure 4-1. Overall pipeline of the proposed automatic splenic AAST grading algorithm. Splenic localization is first performed, which crops irrelevant slices on abdominopelvic CT cranial and caudal to the spleen and neighboring soft tissue that may harbor foci of active bleeding. Active bleeding is detected on portal venous CT scans and pseudoaneurysm is detected on arterial CT scans by Faster RCNN on localized axial sections. SPD is segmented by nn-UNet, and volume is calculated. Active bleeding and pseudoaneurysm detection and SPD volume is then fed into a directed graph for AAST grading prediction.

Table 4-I. Comparison of the 2018 spleen AAST splenic organ injury scale (OIS) grading criteria and lesions in our method selected using available evidence and expert knowledge. SPD- splenic parenchymal disruption; PSA- pseudoaneurysm; AB- active bleeding.

	AAST grading criteria	Lesions in our system
Grade I	Subcapsular hematoma $<$ 10% surface area Parenchymal laceration $<$ 1 cm depth Capsular tear[1]	SPD $<$ 14 mL
Grade II	Subcapsular hematoma[2] 10-50% surface area Intraparenchymal hematoma $<$ 5 cm Parenchymal laceration 1-3 cm	
Grade III	Subcapsular hematoma $>$ 50% surface area Ruptured subcapsular/intraparenchymal hematoma[3] $>=$ 5 cm Parenchymal laceration[4] $>$ 3 cm depth	SPD $>$ 14 mL with no PSA or AB
Grade IV	Any injury in the presence of a contained splenic vascular injury[5] Parenchymal laceration producing $>$ 25% devascularization[6] Active bleeding confined within splenic capsule[7]	PSA
Grade V	Any injury in the presence of a splenic vascular injury with active bleeding extended beyond the spleen into the peritoneum[8]	AB

Note: Capsular tear[1] and subcapsular hematoma[2] are ignored in our simplified system (see discussion). Laceration and intraparenchymal hematoma[3,4] are grouped as splenic parenchymal disruption. Contained splenic vascular injury[5] is synonymous with pseudoaneurysm on CT. In our dataset, only two patients with ground truth consensus grade of IV had laceration with no detected pseudoaneurysm, precluding derivation of a data-driven cut-off differentiating grade III and IV lesions by SPD volume[6]. Current literature does not differentiate between active bleeding confined to or extending beyond the capsule[7,8].

Table 4-II. Ground-truth descriptive statistics of the clinical splenic injury dataset.

	n (%)
Total	174 (100)
Active bleeding	28 (16)
Pseudoaneurysm	54 (31)
PSD $\geq$ 1cm^3	93 (53)
Expert consensus grade	
Grade V	18 (10)
Grade IV	50 (29)
Grade III	15 (9)
Grade I & II	91 (52)

Note: All CT studies in the ABPSA dataset (n = 68) has pseudoaneurysm (n = 31, 46%), active bleed (n = 54, 80%), or both (n = 17, 25%).

Table 4-III. Faster RCNN detection results for active bleeding and pseudoaneurysm.

	AUC	accuracy	sensitivity	specificty	PPV	NPV
Active Bleeding	0.84	88%	82%	91%	91%	83%
Pseudoaneurysm	0.79	83%	91%	78%	94%	71%

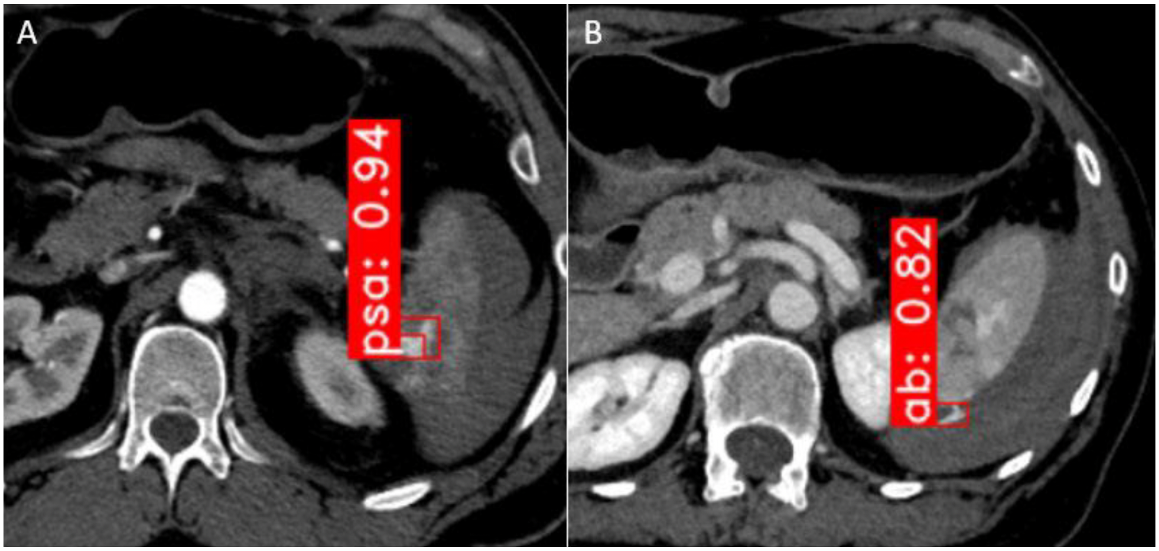


Figure 4-2. pseudoaneurysm (“psa”, part A) and active bleed (“ab”, part B) box detection on arterial and portal venous phase images, respectively. Detection is achieved using Faster R-CNN with a very deep (ResNeXt-101) backbone. Numbers shown refer to probability of correct detection as a fraction of 1.

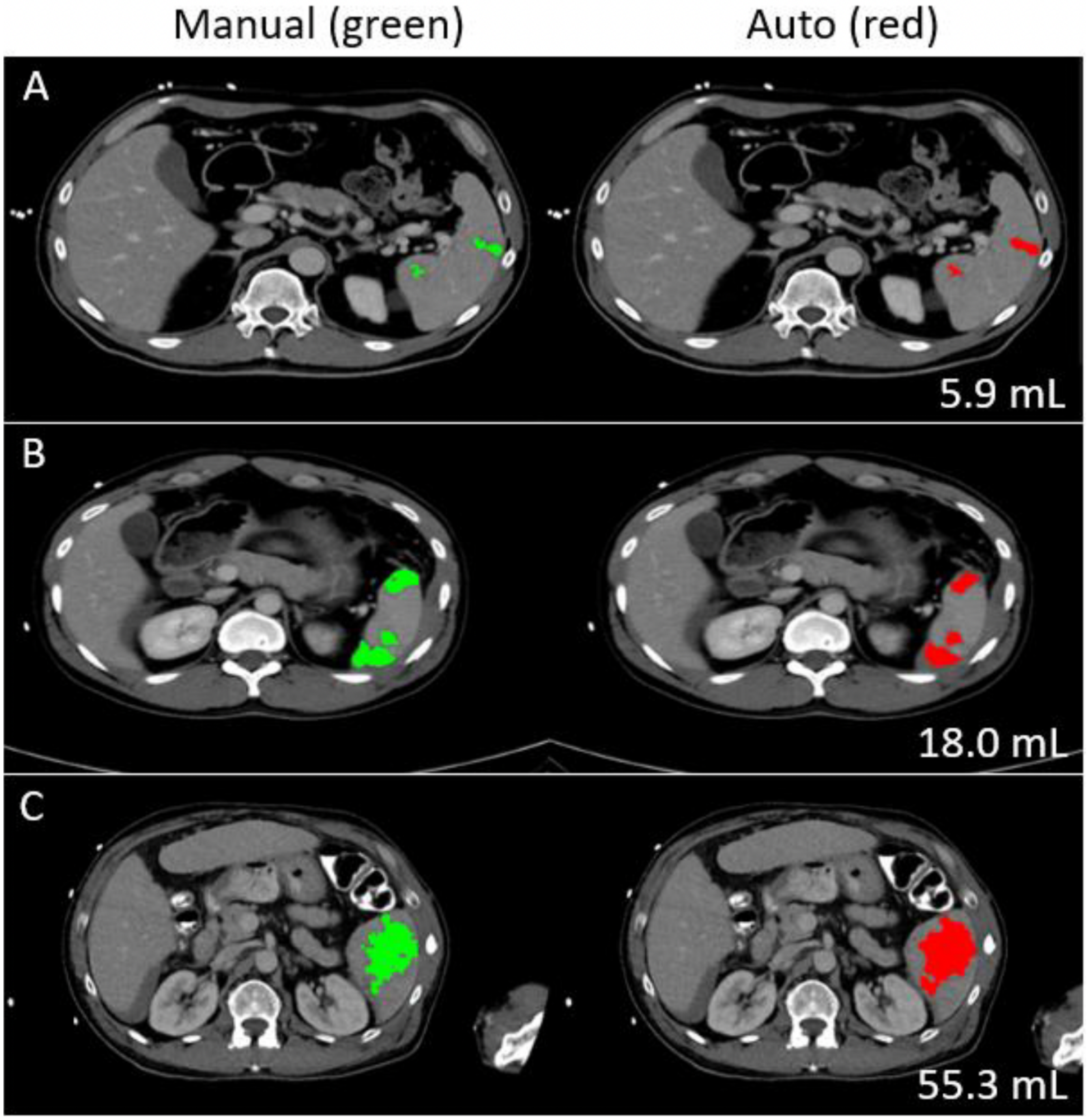


Figure 4-3. Splenic parenchymal disruption (SPD) segmentation/quantitative visualization for a range of volumes. On regression, an optimal cut-off of 14 mL distinguished between low grade (I and II) and grade III lesions.

Chapter 5

An Interactive Approach to Region of Interest Selection in Cytologic Analysis of Uveal Melanoma Based on Unsupervised Clustering

Facilitating quantitative analysis of cytology images of fine needle aspirates of uveal melanoma is important to confirm diagnosis and inform management decisions. Extracting high-quality regions of interest (ROIs) from cytology whole slide images is a critical first step. To the best of our knowledge, we describe the first unsupervised clustering-based method for fine needle aspiration cytopathology images that automatically suggests high-quality ROIs. Our method is integrated in a graphical user interface that allows for interactive refinement of ROI suggestions to tailor analysis to any specific specimen. We show that the proposed approach suggests ROIs that are in very good agreement with expert-extracted regions and demonstrate that interactive refinement results in the extraction of more high-quality regions compared to purely algorithmic extraction alone.

5.1 Clinical Background

Uveal Melanoma (UM) is the most common primary intraocular malignancy in adults [39]. As standard care for UM, Fine Needle Aspiration Biopsy (FNAB) is often performed to confirm the diagnosis and to obtain cell aspirates for both Gene Expression Profile (GEP) and cytopathology image analysis for prognostication. According to recent analysis, primary UM clusters in two distinct subgroups according to its GEP; the first corresponding to low grade melanoma with little to no metastatic risk, and the second corresponding to high grade melanoma with high metastatic risk, which results in 6 times of 5-year probability of metastatic death [254]. While GEP analysis of fine needle aspirates has shown good accuracy for identifying patients at high risk of metastatic disease, the only commercially available test is expensive, requires special storage and transportation, has a long turn around time and is only available in the US. Most importantly, despite its efficacy, the commercial GEP test still occasionally fails resulting in unpleasant clinical surprises and unexpected early metastatic death. There is increasing evidence that the underlying genetic profile affects cancer growth on multiple scales. Radiomics, for example, exploit this observation to develop imaging-derived biomarkers that are informative for prognosis [40]. We hypothesize that such multi-scale analysis will also be useful for prognosis in UM. Specifically in addition to GEP, we would like to extract imaging-features from the cytopathology images. In addition to complementing GEP analysis, such cytology-based test could provide a cheap and widely available alternative for prognostication of UM [255]. However, pathologist analysis of cytopathology images is infeasible, as 1) it is a very time-consuming and tedious task, and 2) none of the manually defined cytopathological features proved particularly robust for predicting metastatic risk. One reason is the cytopathology images exhibit much higher variation in cell quality and artifact compared to histology Whole Slide Image (WSI) which contains an entire slice of tissue.

To reach this goal, we need to facilitate or even automate quantitative analysis of the cytopathology images. Unlike histology WSI where several learning-based algorithms for Region of Interest (ROI) extraction have been proposed [256–259], all existing approaches for high-level cytopathology image analysis operate on manually identified ROIs [260–262]. To this end, we develop an interactive tool that our envision will be beneficial in two ways: First, it can be deployed in pathologist-centric workflows to guide pathologist review, thereby reducing the experts workload. Second, the tool provides an opportunity for pathologists to guide algorithmic evaluation, e. g. by refining the content that is submitted for automated analysis of the slide, e. g. for GEP classification. Such an interactive design may prove beneficial in building trust, accelerating workflows, and reducing mistakes, of both automated algorithms and pathologists.

We present our first steps in this direction that consider the extraction of high-quality ROIs (areas with multiple clear cancer cells) from gigapixel-sized histological architecture, cytopathology images. Further analysis of high-quality ROIs is described in Chapter 6. We propose a Human-Interactive Computationally-Assisted Tool (HICAT) that supports ROI selection with a 2-step coarse-to-fine unsupervised clustering. Coarse-to-fine concepts are widely used due to the small targets (e.g. lesions and organs) in medical imaging. Spatial coarse-to-fine segmentation is applied to target small organs and lesions [263–267]. Spatial coarse-to-fine clustering is also commonly used to extract ROIs from high spatial resolution WSIs and several machine-learning approaches exist for this task [256, 257]. The coarse-to-fine concept in our algorithm aims to resolve the high-quality ROI imbalance problems in cytopathology images. The HICAT system also provides **interactivity** to allow for patient-specific refinement of ROI selection at application time. This refinement provides insight into and some control over the region selection, and results in the extraction of more informative regions compared to the purely algorithmic extraction. Such human-machine partnership may contribute to pathologists building trust in AI-assisted tools. The current research community in human interaction with deep learning so far has been largely limited to segmentation problems [268–272], but our algorithm offers human interaction in a brand new direction for high-quality ROI extraction.HICAT increases Recall in ROIs from 7.44% to 42.32%, while Precision remains the same 83%. On average, 1318 ROIs per cytopathology image are extracted, which contains enough information for further analysis. Our AI-assisted ROI selection workflow is more than 10 times faster than manual ROI extraction by pathologists that was used previously [41].

5.2 Method

Given a cytopathology image, we seek to extract square-shaped ROIs, similar to those shown in Figure 5-2(a), which lend themselves well for further cell-level algorithmic analysis. Our ROI extraction pipeline contains of a 2-step clustering that is followed by an interactive decision boundary definition to assign image-quality to centroids. The clustering algorithm will be discussed in Section 5.2.1 and Section 5.2.2. The first step aims to remove blank images, i.e., Figure 5-2(g), to greatly reduce processing time for the second step, which further clusters the selected ROIs based on image content. After the 2-step clustering, a global decision boundary for all cytopathology images is defined by centroid-level human annotation. Interactive refinement of this decision boundary is then possible for every patient and cytopathology image to improve the algorithmic ROI selection based on centroid annotation (Section 5.2.3).

5.2.1 Step-1 Clustering

The given cytopathology image is first down-sampled such that each pixel in the resulting image corresponds to the average signal within one area. The size of this area is only constrained by its compatibility with the following clustering steps. We found the size 512×512 is able to perform sufficiently well. K-means clustering is then used to cluster pixel intensities into 2 centroids that intuitively correspond to regions with bright and dark average intensities. Since cytopathology images are acquired with the bright-field technique, pixels with low and high intensities correspond to regions with

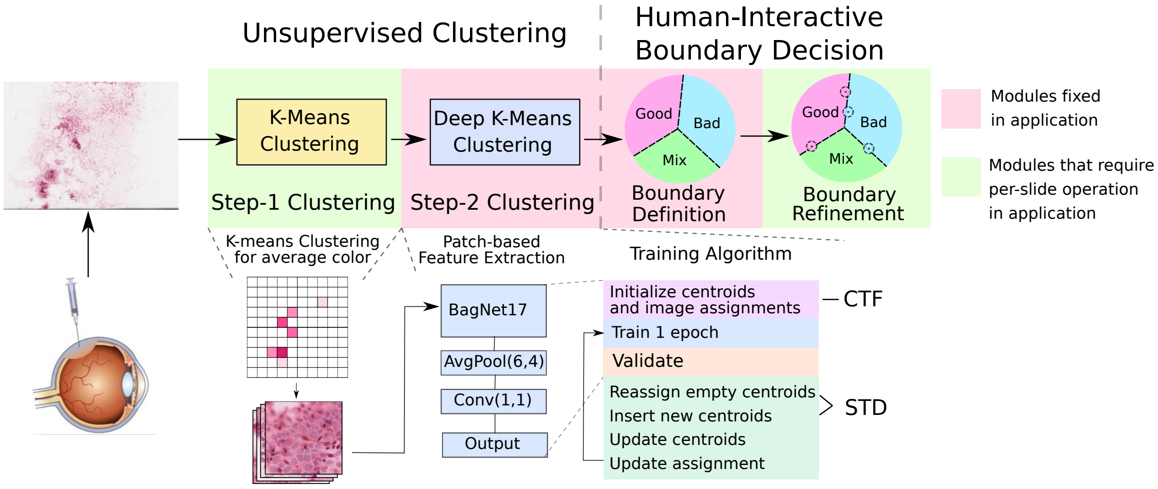


Figure 5-1. Overview of the HICAT.

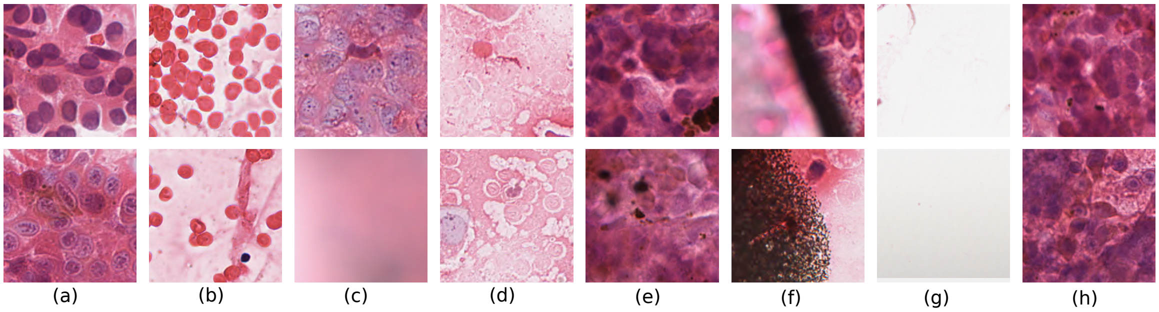


Figure 5-2. Different types of ROIs in cytopathology images. (a) High-quality ROIs, which contain more than 3 clear cancer cells. (b) Blood cell ROIs. (c) Blurred ROIs. (d) Fluid ROIs. (e) Multi-layer cell ROIs. (f) Artifact ROIs. (g) Blank ROIs. (h) Borderline ROIs, which contain more than 3 clear cancer cells, but contains a large portion of low-quality areas.

high and low tissue content, respectively. We select the darker centroid for further processing via Step-2 clustering in Section 5.2.2. Because the exact magnitude of bright and dark centroid intensities varies with cell distribution and illumination, this scheme is applied to every cytopathology image independently.

5.2.2 Step-2 Clustering

Step-2 clustering aims to separate high-quality images with more than 3 clear cancer cells from low-quality images that either show blood cells, multiple layers of cells and fluid, are blurred or otherwise corrupted with artifact. Examples of such images

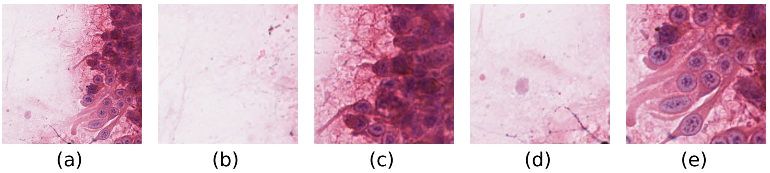
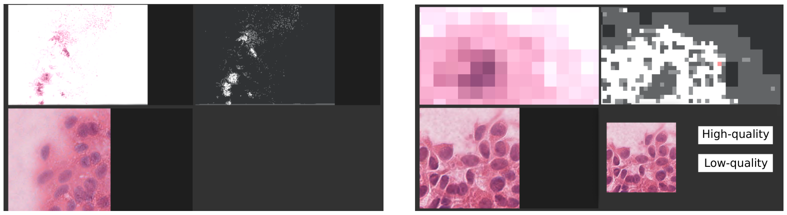


Figure 5-3. An example of Step-1 clustered area and some of the corresponding Step-2 clustering ROIs. (a) Step-1 area. (b) Top-left corner. (c) Top-right corner. (d) Bottom-left corner. (e) Bottom-right corner.

are provided in Figure 5-2. Since this separation is based on image content that, in cytology, can vary considerably across pixels (cf. Figure 5-3), a patch-based network is applied to perform clustering on 228 × 228 pixel ROIs in naive resolution which is much smaller than the areas extracted from Step-1 clustering. These patches are extracted with a stride of 128 from the ROIs selected in Step-1 clustering. A previous state-of-the-art patch-based method, BagNet17 [273] is used as the backbone. The input images of size 512 × 512 pixels are first down-sampled 4 times and an average pooling layer with kernel size 6 and stride 4 is attached after the final residual block, so that each output pixel corresponds to one desired patch (if using other parameters, the receptive field's size and stride cannot be guaranteed to take on the desired value). Finally, a convolutional layer with kernel size 1×1 compresses the feature into a lower-dimension space with dimension d. We follow [274, 275] to involve k-means clustering for the d-dimension network outputs. K-means centroids and patch assignments are initialized by the pre-trained network and are fixed in the training phase. L2 loss is applied to force patch features to be close to the assigned centroid. Centroids and patch assignments are updated during the validation phase. We reassign empty centroids during training to avoid trivial parametrization. Step-2 clustering is trained on all cytopathology images simultaneously.

In order to reduce the number of centroids that focus on fluid and artifact images, we introduce a centroid-based coarse-to-fine clustering strategy. Only a portion of centroids are initialized first, and new centroids are inserted during training in order



(a) (b)

Figure 5-4. Examples for cytopathology image-specific ROI refinement GUI. For each screenshot, top-left image is the down-sampled WSI, top-right image shows the corresponding spatial states for all ROIs, white/light grey/grey means high-/mix-/poor-quality ROIs. Dark grey corresponds to blank images removed by Step-1 clustering. Pink pixels correspond to uncertain ROIs. Bottom left image is the corresponding full resolution ROI that the mouse hovers over. By double clicking the pixel on the down-sampled WSI or state image, a window in bottom right will pop out for annotation. (a) shows the overall behaviour of the state image. (b) shows the zoomed-in version for detail visualization.

to increase the probability of these centroids to account for cell images, which is referred as *CTF* in Figure 5-1. We reassign/insert empty/new centroids around the centroid with the largest standard deviation of its assigned samples in feature space, instead of the centroid with the largest number of samples [274, 275]. It is referred as *STD* in Figure 5-1. This is because of 2 reasons: 1) A considerable number of fluid and artifact images exists and there is no use to further insert centroids for these images. 2) Fluid and artifact images are easier to separate because of the difference in complexity compared to cell images. Consequently, centroids with cell images tend to have larger standard deviation among the assigned samples in feature space, so that inserted/re-assigned centroids are more likely to focus on cell images. The re-assignment and insertion is processed during the validation phase.

5.2.3 Interactive Centroid Assignment and Refinement

After Step-2 clustering, every centroid contains ROIs that exhibit similar appearance. However, at this point it is still unclear which of the ROIs in the centroids are high-/low-quality. To provide this semantic definition with minimal manual annotation requirement, we developed a Graphical User Interface (GUI) that allows for rapid centroid annotation. To this end, 10 ROIs from 10 random centroids are displayed for the user to classify. After several iterations, each centroid has more than 10 high-/poor-quality annotations. The ratio of high-quality ROIs classified to every centroid is then used to define a centroid-level boundary that separates between high- and low-quality ROIs. Because cell quality in cytopathology images has large variation, some ROIs cannot be clearly classified as high-/low-quality, e.g., Figure 5-2(h). Therefore, we allow for some mix-quality centroids that contain roughly an equal number of high-/low-quality ROI annotations. Although there exists high-quality ROIs in mix-quality centroids, we exclude them to avoid introducing poor-quality images to influence further analysis.

During application, due to high variations in cytopathology images, the classifier based on the above procedure may not perform perfectly when suggesting ROIs in new cytopathology images. To allow for the refinement of ROI suggestions, a patient-specific refinement tool is created for pathologists to interact with, as shown in Figure 5-4. Specifically, high-/low-/mix-quality assignments from boundary definition are visualized and synchronized with the corresponding cytopathology image. The user can hover the mouse over the cytopathology image to display the underlying ROI in native resolution, and can simply click it to re-annotate if necessary. In this case, the selected ROI and all ROIs with similar features $\{x, \text{ where } ||x - F||_2 < \lambda_2 L_1\}$ are all re-annotated, where F is the selected ROI's feature, L_1 is the distance to the closest centroid and λ_2 is a constant. Uncertain ROIs are also identified and displayed to users as recommended for re-annotation. Using L_1, L_2 as the distance of an ROI feature to the 2 closest centroids. The ROI is considered uncertain if the two closest centroids are high- and low-quality, respectively, and satisfies

$$\frac{||L_1 - L_2||_2}{\min\{L_1, L_2\}} < \lambda_1 \tag{5.1}$$

where λ_1 is a constant. The result of every click re-annotation is reflected in real time. The user has full control over when to stop the refinement.

5.3 Experiment

5.3.1 Experiment Setup

Dataset: The dataset we use includes 100 cytopathology images from 88 UM patients. The cellular aspirates obtained from cytopathology images of each tumor were submitted to cytology and GEP testing. The cytology specimen was flushed on a standard pathology glass slide, smeared, and stained with hematoxylin and eosin. The specimen submitted for GEP was flushed into a tube containing extraction buffer and submitted for DecisionDx-UM testing. Whole slide scanning was performed for each cytology slide at a magnification of 40x, using the Aperio ScanScope AT machine, and the high-magnification digital image was examined using the Aperio Imagescope software.

516 areas of size 1716×926 are manually extracted and annotated from 20 slides by an expert pathologist. Every area is split into 8 small areas with equal size. Each small area is further split into 9 ROIs where the stride of ROI extraction is half of their width and height. All of these ROIs are annotated as high-/low-quality images, which results in $37,152$ annotated ROIs. The criterion for high-quality images is the same as Figure 5-2(a). All our experiments are trained on the remaining 80 slides and tested on the 20 slides with annotations.

Implementation details: $259,203$ areas are extracted by Step-1 clustering. In Step-2 clustering, each area corresponds to 9 ROIs with size 228×228, which results in a total of $2,332,827$ ROIs for training. The length d of the output feature vector is 16. Centroid-based coarse-to-fine clustering is first initialized with 32 centroids. 4 new controids are inserted after every training epoch until a total of 100 centroids exists. We implement the model using PyTorch [276] for Step-2 clustering, and initialize them with ImageNet pre-trained weights provided by [273]. All models are optimized by Adam [277] with a learning rate of 10^{-3}. All interactive centroid assignments and

Table 5-I. Ablation study for clustering algorithm. DeepCluster (DC) is DCN [274] with BagNet17 [273] as backbone. "CTF" indicates the use of the proposed centroid-based coarse-to-fine strategy. "STD" indicates the use of the proposed mechanism of inserting/reassigning new/empty centroids to be around the centroid with the largest standard deviation of its assigned samples in feature space. (Otherwise, to be around the centroid with most samples). Numbers of high-/low-quality centroids are also reported.

Model	Recall_{gb}	Recall_{gmb}	Precision_{gmb}	Accuracy	#high-quality	#low-quality
DC [273, 274]	11.74%	7.44%	83.17%	61.43%	10	60
DC+STD	34.71%	7.89%	85.99%	63.63%	18	43
DC+STD+CTF	**51.38%**	**27.83%**	**91.56%**	**70.90%**	23	51

specific boundary refinement were performed by an expert pathologist. During centroid definition, centroids with greater than 70% of ROIs annotated as high-quality are classified as high-quality centroids, while centroids with fewer than 30% are classified as low-quality centroids. The other centroids are mix-quality centroids. For boundary refinement, the parameters are $\lambda_1 = 0.2, \lambda_2 = 0.5$.

Evaluation metrics: The final goal for our proposed extraction is to maximize the number of high-quality ROIs and to minimize the number of low-quality ROIs provided for further analysis. To evaluate our success, we calculate the recall, precision and accuracy on the ROIs in the 20 slides with manually extracted ROIs. Because there exist mix-quality centroids, we first report recall and precision for images only in high-/low-quality centroids, denoted as $\text{Recall}_{gb}, \text{Precision}_{gb}$. We also report recall, precision and accuracy for all annotated images, by treating mix-quality centroids as low-quality centroids, denoted as $\text{Recall}_{gmb}, \text{Precision}_{gmb}$ and Accuracy. Because Precision_{gb} is the same as Precision_{gmb}, only Precision_{gmb} is recorded.

5.3.2 Ablation Study for Clustering Algorithm

In order to compare different clustering algorithms, human-interactive boundary definition is performed separately for all models to classify high-/mix-/low-quality centroids by the same expert pathologist. We conduct an ablation study for clustering

Table 5-II. Ablation study for human interactive patient-specific boundary refinement.

Model	Recall$_{gb}$	Recall$_{gmb}$	Precision$_{gmb}$	Accuracy
without Boundary Refinement	51.38%	27.83%	**91.56%**	70.90%
HICAT	**59.47%**	**42.32%**	83.09%	**74.18%**

algorithm to analyze the contributions of its novel components. The baseline is the combination of the deep clustering network, DCN [274], with BagNet17 [273] (referred to as *DeepCluster*) with 100 centroids. The performance by adding the two novel components: centroid-based coarse-to-fine concept (referred to as *CTF*) and the centroid insertion/reassignment algorithm (referred to as *STD*) is compared. The Step-1 clustering is kept the same across all models, which eliminates 96.5% areas as blank areas. Results are summarized in Table 5-I.

The effect of our proposed centroid insertion/reassignment algorithm is reflected in the comparison of *DeepCluster* v.s. *DeepCluster+STD*. Recall$_{gb}$ and Precision$_{gmb}$ increase from 11.74% and 83.17% to 34.71% and 85.99% by using *STD*. Improvements are due to our observation that standard deviation of the assigned samples are efficient to tell apart centroids for high-/low-quality images. More centroids for high-quality images result in better performance.

The effect of centroid-based coarse-to-fine method is reflected in the comparison of *DeepCluster+STD* v.s. *DeepCluster+STD+CTF*. By adding the centroid-based coarse-to-fine module to *DeepCluster+STD*, we observe substantial improvements in Recall$_{gmb}$ and Precision$_{gmb}$ which increase from 7.89% and 85.99% to 27.83% and 91.56%, respectively. The improvement is in line with our motivation and hypothesis that more centroids are assigned to focus on images with different cells and various visual quality. The increase in the number of high-quality centroids further supports our hypothesis.

5.3.3 Ablation Study for Interactive Refinement

The performance of interactive refinement of ROI suggestion is shown in Table 5-II. Cytopathology images and the ROIs' labels after centroid definition are synchronously visualized as Figure 5-4. An expert pathologist finished the human interactive boundary refinement for all testing cytopathology images. Less than 50 re-annotation clicks are performed for each slide. The pathologist stopped the process for each slide, once he determined there were adequate high-quality ROIs selected for further analysis and few low-quality ROIs exist. Comparing with/without boundary refinement shows that Recall$_{gmb}$ goes drastically up from 27.83% to 42.32%. The reduced precision from 91.56% to 83.09% may be attributed to a conservative selection of the pathologist. However, since adequate high-quality ROIs are still available for further analysis, this decrease is likely not problematic. The boost in performance is due to the variation in different cytopathology images. Pathologists may interact with our tool to adjust the inclusion criteria based on a specific cytopathology image, e.g., when few cells are visible, the selection criteria for high-quality ROIs can be relaxed. Finally, 1318 ROIs are extracted on average per cytopathology image, which contain adequate information for further analysis. The whole application process takes 15 minutes per cytopathology image, which is more than 10 times faster than manual ROI extraction. (3 minutes for 2-step clustering and 12 minutes for boundary refinement.)

5.4 Conclusion

In this work, we propose an interactive and computationally-assisted tool for high-quality ROI extraction from cytopathology images. Our method relies on 2-step unsupervised clustering of ROI appearance and content to automatically suggest ROI of acceptable quality. These suggestions can then be refined interactively to adapt ROI selection to specific patients. We hope to contribute effective tools that support quantitative analysis of cytopathology images to, in the future, improve the prognostication of patients suffering from UM.

Chapter 6

Explainable AI and Human-Machine Teaming for Cancer Subtyping from Digital Cytopathology

Algorithmic decision support is rapidly becoming a staple of personalized medicine, especially for high-stakes recommendations in which access to certain information can drastically alter the course of treatment, and thus, patient outcome; a prominent example is radiomics for cancer subtyping. Because in these scenarios the stakes are high, it is desirable for decision systems to not only provide recommendations but also supply explainable reasoning in human-machine teaming support thereof. For learning-based systems, this can be achieved through the explainable design of the inference pipeline. Herein we describe an automated yet explainable system to assist pathologists with cancer subtyping with digital cytopathology images. Following a human-centered design, we first perform formative user research to understand end users' needs and requirements. It reveals that pathologists mainly analyze the cell composition of cytopathology images for cancer subtyping. We strictly follow what pathologists believe to be explainable and informative to design deep learning based models for cancer subtyping by explicitly analyzing cell composition over cytopathology images. We first consistently embed every automatically segmented cell of a candidate cytology image as a point in a 2D manifold with a fixed projection, which enables reasoning about the cell-level composition of the tissue sample, paving the way for explainable subtyping of the biopsy. Finally, a slide-level cell composition analysis is completed with rule-based symbolic reasoning. This process results in a simple rule set evaluated automatically but highly transparent for human verification. On our in house cytopathology dataset of 88 uveal melanoma patients and a public dataset of 60 cervical cancer patients, the proposed method achieves an accuracy of 87.5 % and 93.1 %, respectively, which compares favorably to all competing approaches, including deep "black box" models. We further conduct a user study to assess the human factors of the proposed algorithm, including user willingness and trust in the algorithm. Among all interactive model components, an efficient cell composition inspection tool greatly improves the reliability and effectiveness of human-machine teaming in cancer subtyping.

6.1 Clinical Background

Cancer subtyping is a high-stakes decision-making procedure for the selection of patients that benefit most from specified therapies and the design of novel targeted agents. Cancer classification is largely based on histopathological, cytopathological, and clinical characteristics, which makes it difficult to implement uniformly, as the individual expertise of the clinicians is often a major determinant [278, 279]. Besides, cancer subtyping with microscopy images is a time-consuming task as microscopy images are giga-pixel level images [280]. Furthermore, it is impossible even for highly trained pathologists to derive cancer subtyping for some diseases, i.e. Uveal Melanoma (UM), because no hand-crafted features have been proven to be robust for cancer subtyping. With advances in computer-aided diagnosis (CAD), deep learning has the potential to address the aforementioned bottlenecks. Deep learning-based CAD is much faster than clinical experts in most clinical tasks [38, 46, 151]. Another crucial advantage of deep learning approaches is their ability to learn task-specific salient features and discriminative morphological patterns to diagnose microscopy images in a standardized and objective manner. However, this superiority comes at the cost of explainability. Classic automatic cancer subtyping and analysis in Whole Slide Images (WSIs) is based on multiple small regions extracted from slides, that then need to be aggregated to a single prediction on the slide level. These methods include majority voting, coarse-to-fine techniques [281–284], and multiple instance learning approaches [285, 286]. Most of these techniques consider all image regions equally in the aggregated predictions and ignore the variations in WSI image quality. Such classic machine learning algorithms offer no insights beyond the final recommendation to pathologists, which has been linked to automation bias and over-trust or dis-trust in such systems [43, 44]. A more explainable algorithm design may enable humans to better calibrate their trust in the recommendation, which would be an important feat for high-stakes decision-making such as cancer subtyping.

Recent literature advocates a human-centered design for explainable models in medical image analysis [16]. Multi-disciplinary teaming between designers and clinical stakeholders in model design is highly recommended to increase the likelihood that the designed model is truly explainable to end users in real clinical applications. Teaming with pathologists in cancer subtyping reveals what pathologists believe is clinically important, which should be the fundamental justification of the explainability of the deep learning model to design. In this work, we propose formative user research with pathologists and reveal that pathologists believe cell type composition should be the most salient and informative feature for cancer subtyping with cytopathology images.

Based on the findings in formative user research, we propose a deep learning algorithm that directly analyzes the cell composition of cytopathology images for cancer subtyping and achieves human-machine teaming in the application. However, classic deep learning techniques, such as Convolutional Neural Network (CNN) can hardly learn the global behavior of cell type information in cytopathology images because they are kernel-based feature extractors and mainly focus on local information. Instead, we propose a neural-symbolic model to fully utilize the overwhelming power of CNN in extracting salient features per cell locally and the explainable nature of symbolic reasoning directly with cell composition. In detail, we first automatically extract features per cell with CNN as cell appearance information. Cell-level features are consistently projected into a 2D space as the embedded and informative cell composition. Finally, we apply an explainable rule-based symbolic reasoning for cell appearance composition analysis in the 2D space. The proposed method can also be naively applied to any microscopy image, such as histopathology images. To the best of our knowledge, we are the first to analyze cell composition for cancer subtyping with deep learning techniques.

After developing the proposed method, we further conduct a user study with four pathologists to assess whether the proposed algorithm is indeed interpretable to pathologists. We compare the performance and human factors of pathologists with different levels of assistance: with no assistance, with the assistance of a “black box” model, and with the assistance of the proposed method in cell composition. With the assistance of our proposed method, both pathologists’ accuracy and human factors, i.e. confidence, willingness, and understandability outperform those with the other two levels of assistance, which indicates that the proposed method is truly interpretable to pathologists.

We apply the entire algorithm for cancer subtyping of two diseases: UM and Cervical Cancer (CC). UM is the most common primary intraocular malignancy in adults [39]. As standard care for UM, Fine Needle Aspiration Biopsy (FNAB) is often performed to confirm the diagnosis and enable UM prognostication. To this end, a molecular test, Gene Expression Profile (GEP), is performed and microscopic Cytology of Fine Needle Aspirates images are created from the biopsy. According to a recent study, there exist two subtypes in UM that can be identified based on its GEP: The first subtype exhibits low metastatic risk, while the second subtype has been linked to high metastatic risk. There is a stark contrast in long-term survival between the two classes: the 92-month survival probability in class 1 patients is 95%, versus 31% in class 2 patients [287]. It is evident that access to UM subtype information is critical for the proper management of patients by providing an appropriate recommendation for metastasis surveillance. However, even after 10 years of development, GEP is still only available in the United States. The technique is also expensive and has a long turnaround time. Thus the proposed method aims to predict the GEP subtypes of UM. CC is the second most common malignancy among women [288]. Pap smear is the identified tool for cervical cancer screening but the sensitivity is approximately 50-80% [289]. In this work, we perform a simple task to predict whether each subject has intraepithelial malignancy or not.

6.2 Method

Our goal is to create an interpretable algorithm to assist pathologists for cancer subtyping. We develop the algorithm for UM prognostication specifically. However, developing an interpretable algorithm by us designers alone is highly vulnerable to finally developing an algorithm that is actually not interpretable to pathologists. The reasons are 1) We designers lack the professional skills required for cancer prognostication with whole slide images. 2) We designers have a huge clinical knowledge gap against true pathologists. Thus, the justification of interpretability for the Machine Learning (ML) system determined by ourselves alone is extremely likely to deviate from the knowledge and need of pathologists. As a result, we first performed formative user research (Section 6.2.1) with pathologists to 1) get a comprehensive picture of UM prognostication and 2) understand how they diagnose cytopathology images and which features they focus on. We determine the justification of interpretability based on formative user research. We then build the interpretable algorithm based on the justification of interpretability in Section 6.2.2. Finally, we perform a user study to assess the interpretability with pathologists in Section 6.2.3.

6.2.1 Formative User Research

To avoid determining justification of interpretability with considerable bias against the need of pathologists, we perform formative user research with pathologists for the current stage of UM prognostication and what pathologists believe to be important in cytopathology images for UM prognostication.

We had multiple meetings with two pathologists and one clinical professor for the formative user research. Pathologists taught us how they diagnose cytopathology images hand in hand. The clinical professor introduced the current literature of UM prognostication with GEP as well as cytopathology images. The current stage of UM prognostication and important findings that pathologists focus on are listed below

1. The current gold standard of UM prognostication is GEP classes. There exist no robust manual features in cytopathology images for UM prognostication. As a result, pathologists cannot perform UM prognostication through cytopathology images, making it a "Super Human Task".

2. Pathologists believe that visual information of cells in cytopathology images is adequate to predict UM prognostication. It is still a super human task only because they have not found effective and robust features among the cells.

3. Diagnosis with cytopathology images is a game of numbers. A single cell's visual feature does not represent the overall behavior of cells in the cytopathology images. A macro analysis with cell distribution over the entire cytopathology images is more likely to reveal the UM prognostication status, such as metastatic risk.

According to the findings in formative user research, we aim to build an interpretable ML model to directly analyze the cell distribution of the cytopathology images. Due to the fact that it is a super human task, we also perform a user study to assess the interpretability of our algorithm with pathologists.

6.2.2 Interpretable Cell Appearance Composition Learning

Following the findings in formative user research, we aim to create an interpretable system to analyze UM cytopathology images and reveal GEP subtype based on the overall cell composition of the sample. We construct cell composition by aggregating the cell appearance of each cell in cytopathology images. However, all existing high-quality Region of Interests (ROIs) in cytopathology images are automatically extracted with our previous work [42]. There exist no cell annotations at all. First and foremost, in Section 6.2.2.1, we describe a cost-efficient way of weakly labeling our dataset to enable supervised learning of the cell segmentation network. Features of the segmentation network are aggregated to generate cell-level features that represent cell appearance information for each extracted cell. The overall cell appearance composition of cytopathology images is made up of the appearance of all cells within it. Thus, all cell-level features within cytopathology images are consistently embedded as points in a 2D space and the point distribution in the 2D space is the projected cell appearance composition (Section 6.2.2.2). Finally, we train a rule-based model in the 2D space to directly analyze the cell appearance composition of cytopathology images (Section 6.2.2.3). The system overview is shown in Figure 6-1.

6.2.2.1 Instance Cell Segmentation

In order to analyze cell appearance composition, the first step is to extract all cells in cytopathology images. However, there exist no cell annotations for the high-quality ROIs that are automatically extracted from cytology images by [42], making it hard

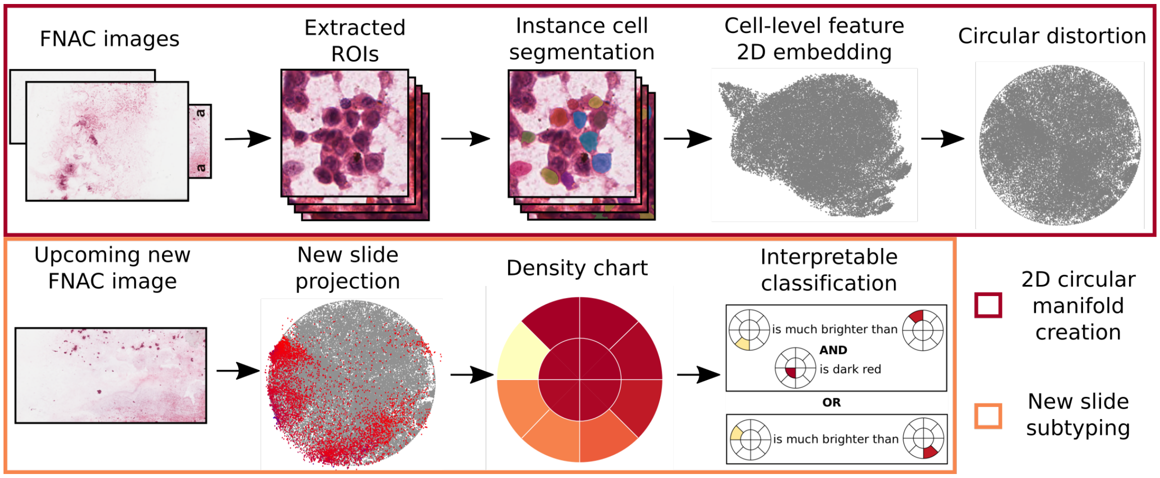


Figure 6-1. System overview of the proposed method. Cell-level features are obtained by aggregation over instance cell segmentation masks and then embedded into a 2D space. Several slides are embedded in this way to create a representative cell appearance space, and the 2D embedding space is subsequently distorted into a circle. For every other cytopathology image, cell representations are extracted and projected with the same embedding process into the circular space, such that one density chart is generated for every slide. Finally, we find an interpretable rule set to classify UM biopsies based on the density charts.

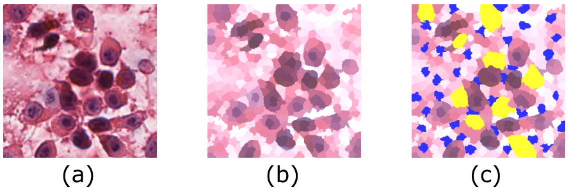


Figure 6-2. The ROI annotation procedure. (a) the extracted high-quality ROI; (b) the generated super-pixels.; (c) the annotations on super-pixels. Yellow and blue regions are annotated super-pixels for cancer cells and background, respectively.

to establish supervised learning for instance level cell segmentation. Thus, we prepare annotations on a small subset with minimal manual labor to enable supervised training of an instance segmentation network. Figure 6-2 presents the annotation procedure. In detail, we randomly select 500 ROIs from the 131k pool and partially annotate super-pixels generated by SLIC [290] to reduce the annotation workload. We group all super-pixels within any cell to prepare instance-level annotations. YOLACT [291] is trained on the partially annotated ROIs, by converting annotated super-pixels into pixel-level annotations. We compute all loss functions, e.g., semantic segmentation loss, only in annotated regions. After we fully trained the cell segmentation network, all ROIs in cytopathology images are fed to the segmentation pipeline to extract cells.

6.2.2.2 Cell-Level Feature Embedding

Previously, pathologists have attempted to quantify different cell components, such as nuclear size and nucleolar size, to predict the behavior of tumors. Our approach is similar to this process but extracts network feature representations of cells, which we assume contain all information about cell appearance. Cell-level features F_c for cell c are extracted from the entire feature map F using the instance segmentation mask M_c with masked average pooling

$$F_c = \mathrm{Avg}(F[M_c]) \tag{6.1}$$

The outputs of the last convolutional layer are used as the feature F for pooling.

To generate the cell appearance composition of cytopathology images, we apply UMAP [292] for all cell-level features to embed cell appearance composition in a 2D space. In order to analyze cell appearance composition, one assumption should be guaranteed that similar cells in different cytopathology images should be close to each other in the embedded 2D space. However, UMAP is a data-driven clustering technique. Applying UMAP individually for each cytopathology image does not hold the assumption. In order to keep the projection parameters as the same, we first generate a “reference” projection by 20 GEP class 1 slides and keep it fixed during application time. All other slides are then embedded with the “reference” UMAP projection, to represent the respective cell composition. We expect slides of distinct GEP classes to have different cell compositions, and thus distribution in the 2D embedding space.

6.2.2.3 Interpretable Classification

Based on our hypothesis that slide-level cell composition, and thus distributions in the 2D cell appearance embedding space, should be different between GEP classes, we

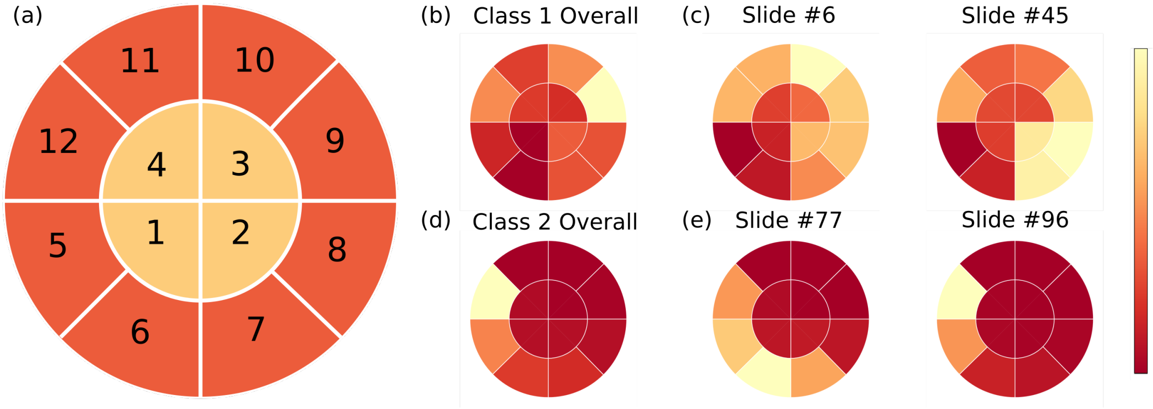


Figure 6-3. (a) The definition of spatial partitioning and density charts in the distorted 2D embedding space. (b) The density chart of all cells in GEP class 1; (c) Two density chart examples of GEP class 1 slides; (d) The density chart of all cells in GEP class 2; (e) Two density chart examples of GEP class 2 slides.

devise an interpretable algorithm that reasons based on these representations. Direct comparisons between distributions, e.g., chi-square test [293] and Kolmogorov-Smirnov tests [294], are complicated and not usually interpretable. Instead, we partition the embedding space and analyze the region densities. To make it easier to define the spatial partitioning of the embedding space, we first distort the space into a unit circle. We treat the center of gravity of all embedded cells as the origin. Then, we normalize to unity the scale of all embedded cells in every degree of angle in polar coordinates, so that the whole embedding space is distorted to a unit circle. Finally, we divide the unit circle equally into 12 regions, as shown in Figure 6-3. Since we posit that each GEP class will have different densities in distinct regions, in addition to the individual densities of these regions (D_i), we define the relative densities (D_i/D_j) as input variables for classification. Finally, an interpretable Bayesian rule set algorithm [295] takes these 78 input variables (12 values, and 66 relations) for GEP classification.

Contrary to other interpretable methods, e.g., logistic regression, the number of input variables will not limit the interpretability of the rule set algorithm, because the number of arguments in each rule can be controlled. In addition, it is different from a random forest (which uses a majority vote) since here, the predicted output

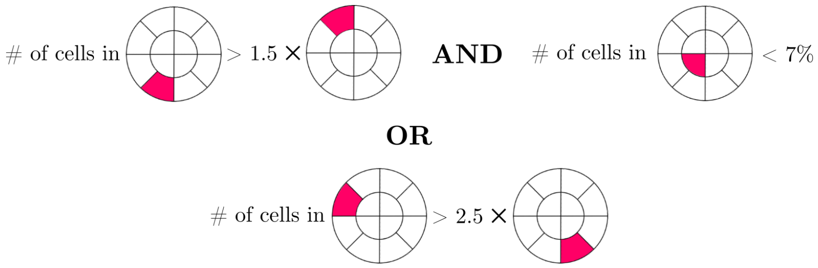


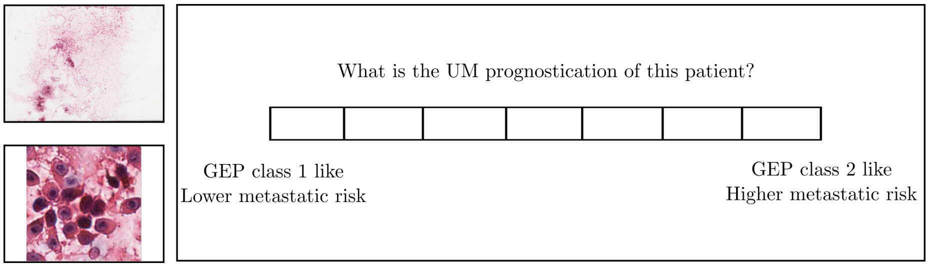
Figure 6-4. Trained rule set for UM dataset.

is positive if the sample obeys at least one rule in the rule set. The rule set for UM cancer subtyping is shown in Figure 6-4.

Datasets The first dataset we use includes 100 cytology samples from 88 UM patients, which refers to as UM dataset. To the best of our knowledge, this is the largest dataset on UM cytology. The dataset contains 50 slides from 43 patients with GEP class 1 and 50 slides from 45 patients with GEP class 2. The cellular aspirates obtained from cytology of each tumor were submitted for cytology and GEP testing. The cytology specimen was flushed on a standard pathology glass slide, smeared, and stained with hematoxylin and eosin. The specimen submitted for GEP was flushed into a tube containing extraction buffer and submitted for DecisionDx-UM testing. Whole slide scanning was performed for each cytology slide at a magnification of 40x. Automatic ROI extraction is performed using [42], resulting in a total of $131,816$ high-quality ROI across all slides.

We also use a second dataset to validate our interpretable cell distribution analysis system, which we refer to as Cervical dataset [296]. The dataset includes total 963 image regions sub-divided into four sets of images representing the four classes of pre-cancerous and cancerous lesions of cervical cancer as per standards under The Bethesda System. The pap smear images were captured in 40x magnification using Leica ICC50 HD microscope which is collected and prepared using the liquid-based cytology technique from 60 patients. In detail, the four classes are "High squamous intra-epithelial lesion", "Low squamous intra-epithelial lesion", "Squamous cell carcinoma" and "Negative for Intraepithelial malignancy" (negative control), which has 9, 4, 4, 43 patients respectively. The number of image regions for each patient ranges from 5 to 37. Due to the extreme imbalance of the number of cases among classes, we combined "High squamous intra-epithelial lesion", "Low squamous intra-epithelial lesion", "Squamous cell carcinoma" as one class named "positive" class and rename "Negative for Intraepithelial malignancy" as "negative" class for training and analysis.

Implementation Details: Super-pixel algorithm Simple Linear Iterative Clustering (SLIC) [290] is implemented following [297]. In SLIC, the number of components is 400, and the Euclidean distance ratio is 1 for UM dataset and the number of components is 200, and the Euclidean distance ratio is 0.1 for Cervical dataset to better fit each dataset. We apply SLIC for 500 image regions in UM dataset and 100 image regions in the Cervical dataset for the manual annotation process. On average, each image region has 9 cells and 38 background super-pixels annotated. The number of prototypes in YOLACT is doubled to 64 to potentially segment more cells within every ROI. The segmentation model is optimized using Adam [298] with a learning rate of 10^{-5} and 4000 iterations with a batch size of 1. We train the model on 80% annotated ROIs and validate on the other 10% ROIs in each dataset separately. During the cell distribution distortion, we empirically split the circular embedding space into 12 partitions for UM dataset, as shown in Figure NEED REFERENCE. We also empirically split the circular embedding space into 54 partitions, which include 18 equal partitions in the inner circle and 36 equal partitions in the outer ring. We observed that the original UMAP cell distribution of the Cervical dataset is too irregular because the size of cells varies extremely in image regions. We further apply θ distortion before ρ distortion. We first sort cell points by θ and distort every 1/360 portion of cell points to fill each 1° degree in θ. These partition settings were found in internal development to yield the best performance compared to other split approaches.

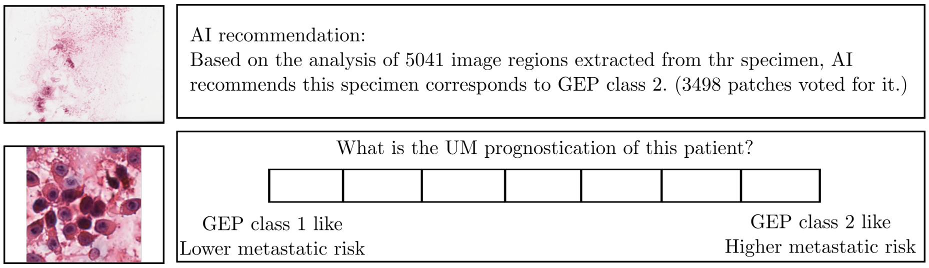


What is the UM prognostication of this patient?

GEP class 1 like
Lower metastatic risk

GEP class 2 like
Higher metastatic risk

Figure 6-5. Prototype of task 1 in the user study. Users need to predict cancer subtyping without any AI assistance.



AI recommendation:
Based on the analysis of 5041 image regions extracted from thr specimen, AI recommends this specimen corresponds to GEP class 2. (3498 patches voted for it.)

What is the UM prognostication of this patient?

GEP class 1 like
Lower metastatic risk

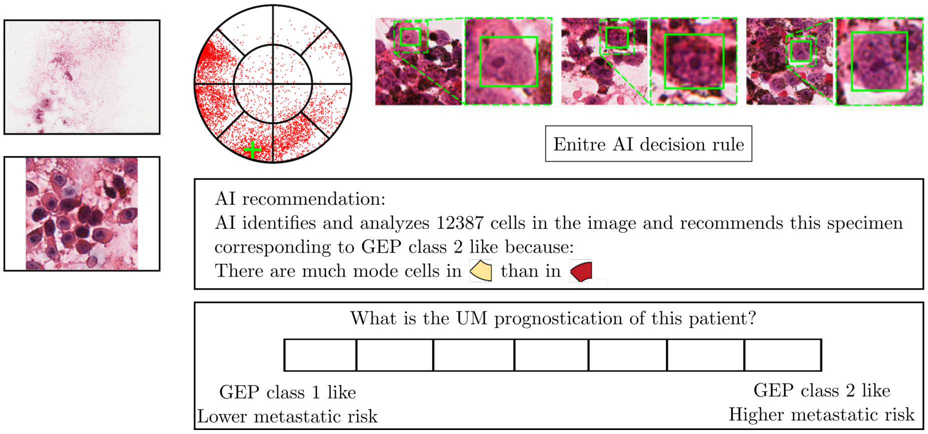
GEP class 2 like
Higher metastatic risk

Figure 6-6. Prototype of task 2 in the user study. Users need to predict cancer subtyping with AI predictions of ROIs by majority votes of predictions in image regions.

We used the cell distribution of 20 slides in GEP class 1 in UM dataset to generate cell projections for application. We also used 10 patients in the negative class in Cervical dataset to generate the cell projection in application time. In application time, all cells that map outside the circular embedding space are projected to the nearest region. For the interpretable classification, we use 80% of the projected slides in both all classes for training (64 for UM; 40 for Cervical) and the other 20% for testing (16 for UM; 10 for Cervical). The rule set algorithm is trained with a simulated annealing procedure as described in [295]. The maximal length of each rule in the rule set is set to 2 to preserve its intelligibility.

6.2.3 User Study

After the development of our interpretable algorithm for UM prognostication with cytopathology images, we further perform a user study to assess whether the algorithm is truly interpretable to pathologists. Instead of constructing the user interface for



AI recommendation:
AI identifies and analyzes 12387 cells in the image and recommends this specimen corresponding to GEP class 2 like because:
There are much mode cells in 🟨 than in 🟥

What is the UM prognostication of this patient?

□ □ □ □ □ □ □

Figure 6-7. Prototype of task 3 in the user study. Users need to predict cancer subtyping with AI predictions based on cell appearance composition. Cell composition is visualized in a pie chart. Users can inspect the cell appearance of any area in the pie chart.

the user study on our own, we first shared a prototype of the user interface for the experts to collect feedback. We iteratively refine the prototype until the experts were satisfied with the prototype.

Our user study consists 3 tasks for pathologists to prognose UM based on cytopathology images with:

1. No ML predictions.
2. ML predictions with an explanation at the small image region level.
3. ML predictions with an explanation about cell distribution.

We created prototypes for each task. They are shown in Figure 6-5, Figure 6-6 and Figure 6-7. Because we planned to perform user study through the internet, the left-hand side of Figure 6-5, 6-6 and 6-7 is an inspection tool for cytopathology images we will develop, which is the same for all tasks. The cytopathology image will be shown in the top left window and users can move, zoom in and zoom out the cytopathology image as other software tools such as QuPath. However, due to the limit of network speed, we only show a thumbnail of the cytopathology image in the top left window. Users can click on the area of interest in the thumbnail image and the corresponding cytopathology image region with full resolution will be shown in the bottom left window. Another part of the prototype that is the same for all tasks is the diagnosis panel. Because UM prognostication with cytopathology images is a super human task, we offer 7 levels of choices for user diagnosis, from very certain that the specimen has a low metastatic risk to very certain that the specimen has high metastatic risk. The middle choice indicates that the user cannot tell the metastatic risk with diagnosis.

In task 1 (Figure 6-5), users directly make the diagnosis with the inspection tool. In task 2 (Figure 6-6), AI recommendations for prognostication: GEP class 1 like or GEP class 2 like is shown to users, together with the voting results in image region level. Users can choose to consider/not consider the AI recommendation on their own in the final diagnosis. In task 3 (Figure 6-7), which is the user study for our proposed method, we additionally show the pie chart that consists of all extracted cells. Users can interact and inspect the cells clustered within the pie chart by clicking the interesting area in the pie chart. The full resolution of the image region together with the bounding box and the zoomed-in version of the image region is shown on the right. Because AI in this task directly analyzes the cell distribution, the AI recommendation switches to provide evidence from the cell distribution according to the trained rule set. The entire rule set will be shown when users click the corresponding button. As in task 2, in task 3, Users are also free to consider/not consider the AI recommendation on their own in the final diagnosis.

The prototype is shown to one clinical expert and 5 Ph.D. students for suggestions. All of them were satisfied with the overall design of the prototype, including the inspection tool of cytopathology images and the diagnosis panel. However, all of them thought the rule set criterion in the prototype is too hard to follow. We modified iteratively the representation of partitions in the pie chart from partition indexes to partition's shape. It showed that the partition's shape can help users quickly find the corresponding partition. The clinical expert also mentioned that showing cell type information within the pie chart would be helpful. Unfortunately, because we do not have cell type annotations and predictions, this may be future work. The clinical expert also mentioned that using the same specimen would make the study across tasks comparable. We applied this suggestion and cropped the cytopathology images differently for each task so that users cannot easily recognize these specimens are the same in different tasks.

We developed the user study according to the refined prototype. We use 6 cytopathology images in each task of the user study. The user interface for tasks 1, 2, and 3 is shown in Figure 6-8, 6-9 and 6-10. Before using the interface, the users are provided detailed instructions on how to use the interface for each task. The user interface of tasks 1 and 2 is the same as the prototype. In the user interface of task 3, when users hover the mouse over each partition of the pie chart, the portion of cells in the corresponding partition is shown beside the mouse icon. We also re-designed the representation of partitions in the AI recommendation. The partitions are highlighted in red in the pie chart and the portion of cells belonging to the partition is also shown above the partition to make it easy for users to compare. In the AI recommendation, the reason for AI prediction is concisely included in the paragraph.

During the user study, we also ask for user feedback per specimen and task by implementing a questionnaire within the web page. Users can select their answer to feedback from 5 choices, from "strongly disagree" to "strongly agree". The middle choice is "fair". The user diagnosis and feedback are recorded for further analysis. For each task per specimen, we ask feedback of

- *you are confident with your diagnosis.*

For tasks 2 and 3, we additionally ask feedback about AI recommendation

- *AI recommendation is easy to understand.*
- *You agree with the AI recommendation.*

- *You considered the AI recommendation in your final decision.*

We also ask for feedback once users complete each task. After completing task 1, we also ask for user feedback of

- *This workflow is easy to use.*
- *You are willing to use this workflow in real clinical practice.*

After completing task 2, we additionally ask for feedback of

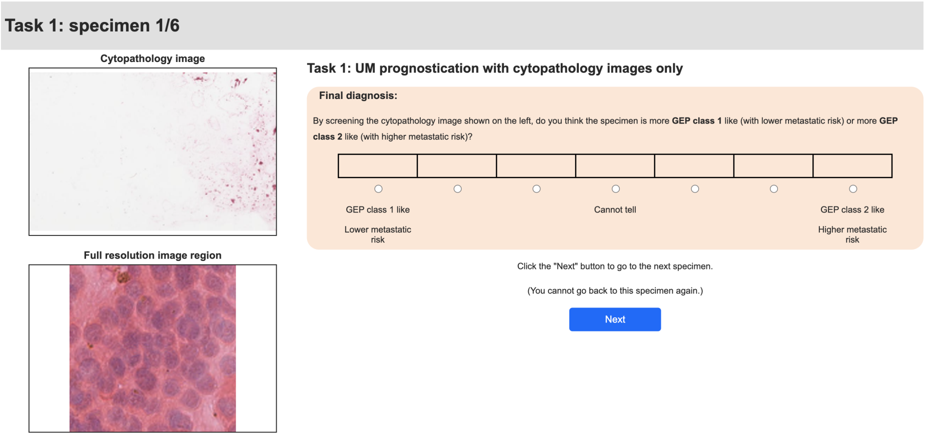
- *AI recommendation accelerates diagnosis.*

After completing task 3, we additionally ask for feedback of

- *Pie chart is efficient to extract cell information.*

User study setup

We performed the user study mentioned in Section 6.2.3 with 4 pathologists. The user interface was developed by Flask [299] and software ngrok offered remote access to the user interface. We recorded the user study with each participant through zoom and we have got the agreement to record from all participants. Each task in the user study contains 6 cytopathology images. The 6 cytopathology images are the same to keep results comparable. Three of them are of GEP class 1 and the other three are of GEP class 2. In each class, one specimen contains few cells and the two others contain numerous cells. In each class, GEP of two specimens are correctly classified by the rule-based algorithm and the other one is miss classified, so the accuracy of the AI in the user study is 66.7%. However, each miss classified specimen is proved to be a clinical surprise, which does not follow typical GEP class survival status. For example, one GEP class 1 cytopathology image is miss classified as GEP class 2 but the patient only survives 2.3 years. One GEP class 2 cytopathology image is miss classified as



Task 1: specimen 1/6

Cytopathology image

Full resolution image region

Task 1: UM prognostication with cytopathology images only

Final diagnosis:

By screening the cytopathology image shown on the left, do you think the specimen is more **GEP class 1** like (with lower metastatic risk) or more **GEP class 2** like (with higher metastatic risk)?

○ ○ ○ ○ ○ ○ ○

GEP class 1 like — Lower metastatic risk

Cannot tell

GEP class 2 like — Higher metastatic risk

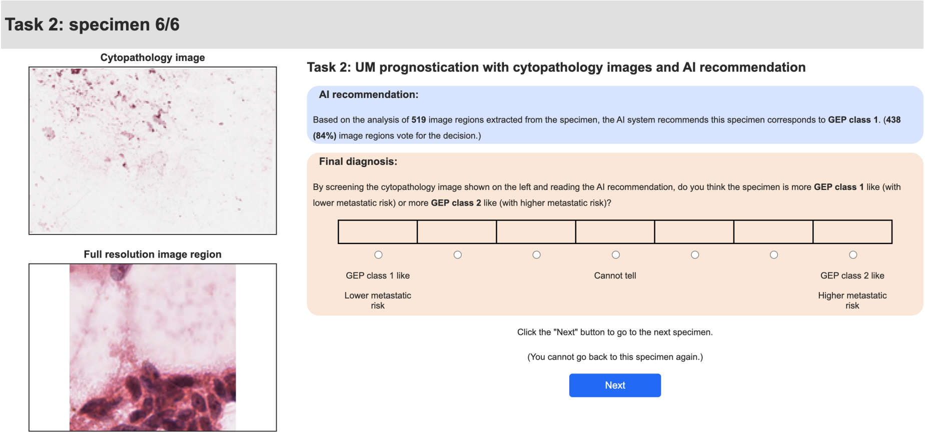
Click the "Next" button to go to the next specimen.

(You cannot go back to this specimen again.)

Figure 6-8. User interface of task 1. Users need to predict cancer subtyping without any AI assistance. Users can move and zoom-in and -out the top left cytopathology image to select the region they are interested in. The selected region is shown in the bottom left image with full resolution.

GEP class 1 but the patient was alive in the last follow-up and has survived at least 6.6 years. As the result, model prediction can be treated as the truth of survival status. The specimens are cropped differently and randomly ordered in each task to avoid pathologists recognizing them in the following tasks. To crop the cytopathology images, we guarantee that cropped specimens still contain a large portion of cells in the original slide.

All participants are asked to prognose UM for each image in every task. There contains 7 slots for users to choose to diagnose GEP classes / metastatic risk for each cytopathology image. We define the first 3 as the participants predicting GEP class 1 and the last 3 as the participants predicting GEP class 2. The middle one is considered as "cannot tell". We use the three defined classes to calculate the accuracy of user diagnosis. All questions for user feedback after each specimen and task have 5 slots for users to choose from, which range from “Strongly Disagree”, “Disagree”, “Fair”, “Agree” and “Strongly Agree”.



Task 2: specimen 6/6

Task 2: UM prognostication with cytopathology images and AI recommendation

AI recommendation:

Based on the analysis of **519** image regions extracted from the specimen, the AI system recommends this specimen corresponds to **GEP class 1**. (**438 (84%)** image regions vote for the decision.)

Final diagnosis:

By screening the cytopathology image shown on the left and reading the AI recommendation, do you think the specimen is more **GEP class 1** like (with lower metastatic risk) or more **GEP class 2** like (with higher metastatic risk)?

○ ○ ○ ○ ○ ○ ○

GEP class 1 like
Lower metastatic risk — Cannot tell — GEP class 2 like
Higher metastatic risk

Click the "Next" button to go to the next specimen.

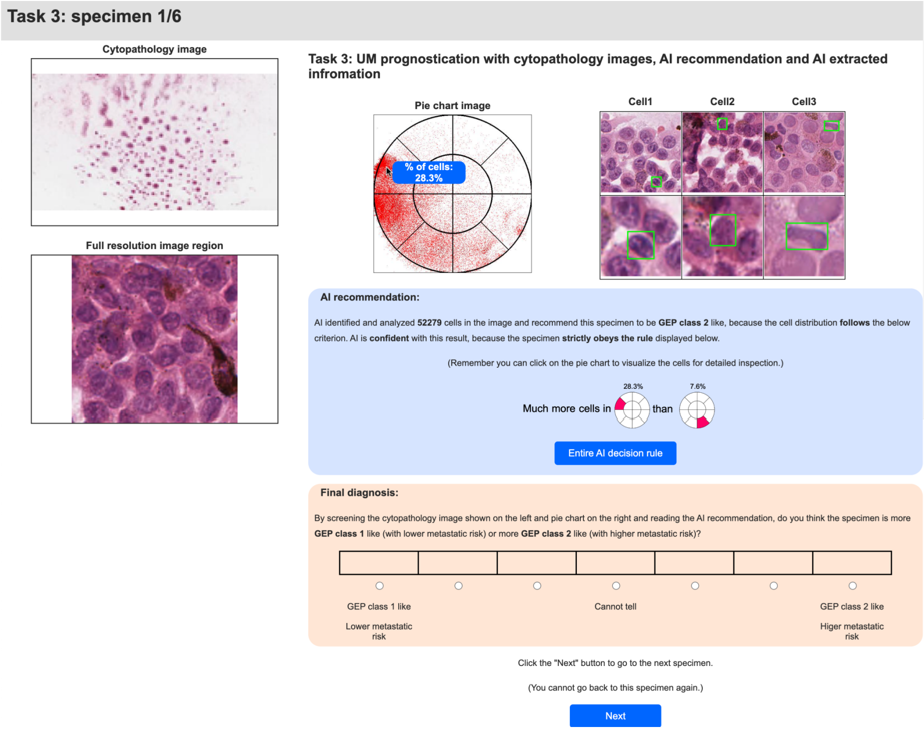
(You cannot go back to this specimen again.)

Figure 6-9. User interface of task 2. Users need to predict cancer subtyping with the assistance of AI predictions with ROIs. The AI predicted class and how many ROIs vote for the decision is also displayed.

6.3 Results and Findings

6.3.1 Cell Appearance Composition Classification

We compare our proposed method on UM cancer subtyping with a previously proposed deep black box model [41, 300] evaluated on the same dataset, which classifies UM subtype directly from ROIs. In [300], slide-level subtype prediction is obtained by simply averaging class predictions for all corresponding ROIs. Both, the black box and our proposed method use the same backbone network architecture (ResNet-50 [301]) and the same training and testing split for a fair comparison. We observe that the accuracy performance of our method (**87.5%**) compares favorably to the black box approach following [300] (83.3%). More importantly, our proposed method is highly likely to be interpretable because the trained rule set is simple to understand, and our method prognoses cytopathology images directly by cell appearance composition analysis which follows pathologists' clinical knowledge that is confirmed in the formative user research with pathologists (Section 6.2.1). There only exist 3 arguments in the rule set, which makes algorithmic recommendations transparent and verifiable, while enabling users to understand overall cell composition. We also evaluate the performance



Task 3: specimen 1/6

Task 3: UM prognostication with cytopathology images, AI recommendation and AI extracted infromation

AI recommendation:

AI identified and analyzed **52279** cells in the image and recommend this specimen to be **GEP class 2** like, because the cell distribution **follows** the below criterion. AI is **confident** with this result, because the specimen **strictly obeys the rule** displayed below.

(Remember you can click on the pie chart to visualize the cells for detailed inspection.)

Much more cells in than

Final diagnosis:

By screening the cytopathology image shown on the left and pie chart on the right and reading the AI recommendation, do you think the specimen is more **GEP class 1** like (with lower metastatic risk) or more **GEP class 2** like (with higher metastatic risk)?

○ ○ ○ ○ ○ ○ ○

GEP class 1 like — Lower metastatic risk; Cannot tell; GEP class 2 like — Higer metastatic risk

Click the "Next" button to go to the next specimen.

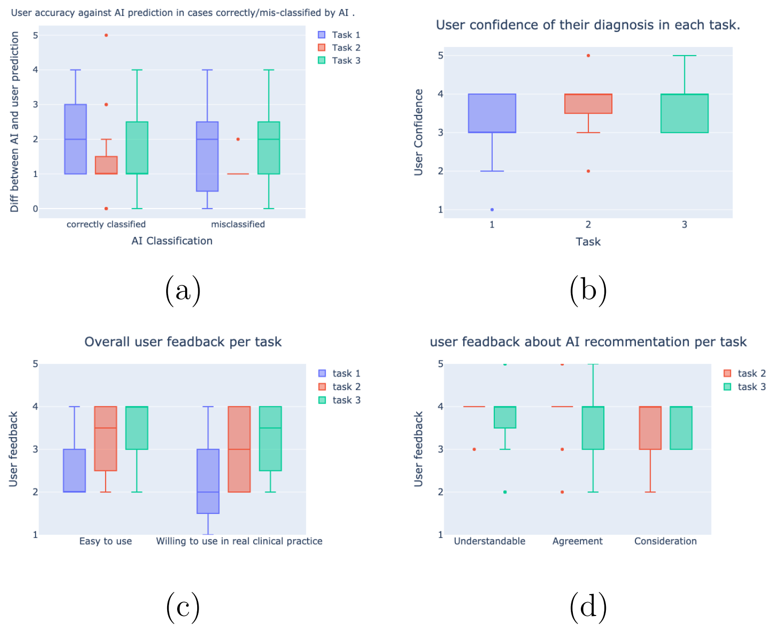
(You cannot go back to this specimen again.)

Figure 6-10. User interface of task 3. Users need to predict cancer subtyping with the assistance of AI predictions based on cell appearance composition. Cell composition is visualized in a pie chart. Users can click any area in the pie chart to inspect the corresponding cell appearance. The AI prediction is shown with the evidence from the rule set criterion.

of rule-based cell appearance composition classification for Cervical cancer subtyping. The accuracy of the trained rule set reaches 93.1%. As a result, our proposed rule-based method analyzing cell appearance distribution achieves overwhelming performance in both UM and Cervical cytopathology images and in both normal v.s. abnormal and low v.s. high metastatic risk predictions.

6.3.2 User Study Findings

We perform a user study to evaluate the human factors of our proposed model and a black box model on UM cancer subtyping. Details of user study design and protocol are shown in Section 6.2.3. Pathologists are asked to perform the same task with 3 different levels of assistance: without any AI assistance; with the assistance of



(a)

(b)

(c)

(d)

Figure 6-11. User study analysis.

the black box model, and with the assistance of the proposed model. We call tasks with 3 different levels of assistance as task 1, task 2, and task 3. We summarized all the results from the user study in the following paragraphs. We further provide corresponding recommendations to pathologists and system designers for participating and designing human-machine teaming in UM prognostication below:

- Pathologists should follow AI predictions in cancer subtyping with cytopathology images to achieve higher performance.
- Formative user research is essential for designers to understand pathologists' needs in cancer subtyping.
- Efficient cell inspection tools and straightforward cell composition visualization are two key successes in human-machine teaming in cancer subtyping with microscopy images.

Human-machine teaming performance cannot exceed machine-alone performance.

Pathologists made their diagnosis for each specimen among *"low metastatic risk (GEP class 1)"*, *"high metastatic risk (GEP class 2"* and *" cannot tell"*. *"Ccannot tell"* is treated as misclassification. The user accuracy in predicting GEP in tasks 1, 2, and 3 is 0.458, 0.542, and 0.542, respectively. The performance of task 1 is the user accuracy without any AI assistance, which is slightly lower than those with AI assistance (tasks 2 and 3). However, the accuracy in all the tasks is much lower than the AI-alone performance (0.667), indicating that *human-machine teaming cannot achieve higher performance than AI alone.*

Pathologists are predicting metastatic risk instead of cancer subtype.

The users' dignosis alignment with AI predictions in tasks 1, 2, and 3 is 0.667, 0.875, and 0.75, which is much higher than those aligned with GEP (0.458, 0.542, and 0.542). Due to the fact that UM prognostication is a super human task and AI prediction perfectly corresponds to survival status, it is reasonable to believe that pathologists are actually predicting metastatic risks of specimens instead of their GEP classes in this user study. With AI recommendations in tasks 2 and 3, pathologists significantly achieve higher accuracy in human-machine teaming according to the t-test ($\mathrm{p}_{\mathrm{value}} = 0.026$). In addition to the three diagnosis classes, pathologists were also asked to rate their confidence in their diagnosis. In detail, rating "1" indicates the pathologist is confident that the specimen is *"GEP class 1"*, and rating "7" indicates the pathologist is confident that the specimen is *"GEP class 2"*. Rating "4" indicates "cannot tell" and the pathologist has no confidence in either *"GEP class 1"* and *"GEP class 2"*. We calculate the difference between users' diagnosis and AI predictions and the statistics of difference is shown in Figure 6-11 (a). Here the difference is calculated by $|u_i - A_i|$, where u_i is the user's diagnosis ratings which range from 1 to 7; $A_i = 1$ when AI predicts GEP class 1 and $A_i = 7$ when AI predicts GEP class 2. There is no significant difference in user behaviors between AI correctly and misclassified cases among all tasks ($\mathrm{P}_{\mathrm{values}}$ for the t-test in task 1, task 2, and task 3 are 0.558, 0.604, and

0.732). It indicates that *users do not have a complementary opinion against AI in UM prognostication.* As a result, pathologists cannot improve human-machine teaming performance. To achieve the highest possible accuracy in super human tasks, the best strategy for pathologists is to follow AI predictions.

Pathologists are more confident with diagnoses made with AI assistance.

We asked pathologists to rate their confidence of diagnosis in each specimen in each task. The levels of ratings are 1 ("*Strongly disagree*"), 2 ("*Disagree*")3 ("*Fair*"), 4 ("*Agree*") and 5 ("*Strongly agree*"). Figure 6-11 (b) shows the user's confidence in each task. Most of the user's feedback about confidence is between 3 ("*Fair*") and 4 ("*Agree*") and the difference between tasks is limited. The pathologists' confidence in task 1, task 2, and task 3 is 3.21 ± 0.74, 3.75 ± 0.65, and 3.79 ± 0.62 respectively. However, user confidence with AI assistance is significantly larger than that without AI assistance. With non-parametric Friedman's test, the user confidence in the three tasks is significantly not the same ($\mathrm{p}_{\text{value}} = 0.00017$). Using a non-parametric Wilcoxon test to further make comparisons between every two tasks, we found that users are more confident with their diagnosis with AI assistance (p-value = 0.002 for task 1 v.s. task 2 and p-value = 0.003 for task 1 v.s. task 3). AI assistance is highly reliable to one pathologist:

> *In those cases that the quality of the slides is poor, then I suppose I would more likely be inclined to use AI as an additional kind of answerary mechanism for interpretation.*
>
> *If AI is confident in the cases, I will use it in practice. If you can tell me AI has confidence over 90%, then I will trust it.*

However, the cell composition explanation does not further significantly increase user confidence compared to black-box models (p-value = 0.655 for task 2 v.s. task 3).

Pathologists are more willing to use AI-assisted systems.

We performed a post-task survey after each task to receive feedback from pathologists on whether the user interface of each task is easy to use and whether they are willing to use the tool in each task in real practice. Figure 6-11 (c) shows user feedback on their interaction with the system. There appears an obvious trend that AI assistance is easy to use and users are more willing to use systems with AI assistance. In detail, the willingness with AI assistance is never lower than that without AI assistance in all specimens.

Pathologists have similar feedback on different AI recommendations.

We further performed a post-diagnosis survey after each specimen's diagnosis of each task to receive feedback from pathologists on whether the AI recommendation is understandable; whether they agree with the AI recommendation; and whether they consider AI recommendations in their final diagnosis. Figure 6-11 (d) shows user feedback on AI recommendations. It is shown that AI recommendations are equally understandable in tasks 2 and 3. Pathologists agree with and use AI recommendations as a reference opinion equally in tasks 2 and 3.

Cell-oriented visualization greatly improves the effectiveness of human-machine teaming

Understanding users' preferences for one interface over the other is of pivotal importance to analyze their impressions. We asked the participants to give us answers about 1) what they liked the most about the interface they had just used, and 2) what challenges they see in using this workflow for UM prognostication. All pathologists prefer the AI-assisted interface to the first interface with no assistance. More importantly, pathologists all expressed their likeness to the cell inspection tool (pie chart) and cell composition visualization. It is much easier to inspect cells with the interaction of the pie chart. The clustering of cell characteristics also helps pathologists to inspect similar cells more efficiently. In overall user behavior,three of the four participating pathologists spent 99% of diagnosis time interacting with the pie chart for cell inspection. The other pathologist also spent > 60% of the diagnosis time with the pie chart and he explicitly mentioned her habit of inspecting cells directly with cytopathology images. All four pathologists also agree on the effectiveness of cell-oriented visualization:

> *I like the pie chart because I like the ability to see cells quickly without trying to search for them.*
>
> *I like the clustering so you see the features of the similar cells.*
>
> *It is definitely easier to find cells in the pie chart.*

Moreover, pathologists expect more clinically-relevant explanations through the pie chart, i.e. clustering cells according to clinically meaningful groups, such as cell types. Surprisingly, they also stated the desired features they want in the pie chart.

> *In real practice, it would be nice for the pie chart to show why and which cells are in the pie chart.*
>
> *Looking at the microscope and having an AI suggestion with specific cells with some kind of annotation will definitely be helpful, some kind of explanatory pie chart.*

6.3.3 Ablation Study

We conduct an ablation study of the rule-based interpretable classification to benchmark its performance against other classification methods, i.e., logistic regression, Support Vector Machine (SVM) and Artificial Neural Network (ANN). We also compare different embeddings, by creating the initial UMAP embedding space with either, GEP class 1 or GEP class 2 slides. The quantitative results are shown in Table 6-I. We also performed the same experiments for Cervical dataset except that we only initialize embedding space by negative classes and results are shown in Table 6-II.

After the embedding space creation, only 80 slides of UM dataset and 50 patients of Cervical dataset remain to train and evaluate the classification models. Therefore, we also introduce an ensemble method to enrich the input data by creating synthetic cell compositions. To create a synthetic slide/patient, we randomly selected 30% cells from one slide/patient and 1% cells from all the other slides/patients in the same class as all the cells in the synthetic slide/patient. Then, the synthetic slide/patient will represent the main pattern of one observed slide/patient but also introduce other variations. We created 100 synthetic slides/patients for each class using this approach, which is indicated as "Ensemble" in Table 6-I and 6-II. The simple ANN we used is fc(8) + ReLU + fc(1), where n in fc(n) means the number of output channels. To evaluate the methods, we perform 100 different and random training/testing splits of our dataset on the patient level and train all models on every split. The mean results and the corresponding confidence intervals are summarized in Table 6-I.

In UM dataset, logistic regression results in the lowest testing accuracy (75.14%) and the rule set achieves the highest performance (87.50%), which is comparable to SVM (82.07%) and ANN (83.71%). Creating the embedding from distinct GEP classes results in similar accuracy of the rule set algorithm (87.50% v.s. 84.33%). As in the previous comparisons to black-box models, the rule set approach has the added benefit of being interpretable. Logistic regression and SVM models suffer in this regard due to the high dimensionality of the input representation (78). Dimensionality reduction techniques, e.g., principal component analysis (PCA), exist but are not applicable here because the number of input variables (78) is larger than the number of training samples (64). Finally, all models reach higher accuracy with the ensemble except SVM. In Cervical dataset, a similar pattern is observed. The trained rule set achieves the highest performance (93.1%) which outperforms all the other methods. However, creating the embedding from distinct GEP classes results in significantly higher accuracy of the rule set algorithm (93.1% v.s. 85.5%). If is not surprising

Table 6-I. Ablation study of interpretable classification with different methods and an ensemble technique for UM dataset. LR refers to logistic regression. Rule Set (class k, $k = 1, 2$) refers to results using the embedding created from class k slides.

	w/o Ensemble		w/ Ensemble	
	Accuracy	# of rules	Accuracy	# of rules
LR	$67.50 \pm 5.56\%$	N/A	$75.14 \pm 9.00\%$	N/A
SVM	$83.00 \pm 6.37\%$	N/A	$82.07 \pm 8.23\%$	N/A
ANN	$82.86 \pm 8.33\%$	N/A	$83.71 \pm 10.15\%$	N/A
Rule Set (class 1)	$\mathbf{86.36 \pm 10.25\%}$	2.28 ± 0.57	$\mathbf{87.50 \pm 9.56\%}$	2.11 ± 0.37
Rule Set (class 2)	$81.93 \pm 8.02\%$	2.06 ± 0.49	$84.33 \pm 10.68\%$	1.96 ± 0.31

because the number of image regions for each subject in Cervical dataset is much smaller than UM dataset, thus a much smaller number of cells for each subject. Due to the Law of Large Numbers, the synthetic subjects created by our ensemble method have larger variance in Cervical dataset, and equally, larger “observed” space in all machine learning algorithms. As a result, models trained with ensemble in Cervical dataset are more generalized than those in UM dataset.

Due to the fact that our segmentation model is not perfect, we also evaluate the rule set model for different segmentation results in UM dataset. We observe that during early training, the segmentation model will first identify the clearest cancer cells, but along with plenty of false positives. As the optimization progresses, fewer cancer cells are segmented but meanwhile much fewer false positives occur. The accuracy of the rule set algorithm for segmentation results after 2000, 3000, and 4000 training iterations is $77.23 \pm 10.98\%$, $84.64 \pm 10.46\%$ and $87.50 \pm 9.56\%$, which suggests that the algorithm favors the output of a highly specific cell segmentation algorithm.

6.3.4 Cell Segmentation Performance

We use Mean Average Precision (mAP) given Intersection over Union (IoU) threshold as the main metric to evaluate cell segmentation performance. The quantitative results

Table 6-II. Ablation study of interpretable classification with different methods and an ensemble technique for Cervical dataset.

	w/o Ensemble		w/ Ensemble	
	Accuracy	# of rules	Accuracy	# of rules
LR	$82.9 \pm 11.8\%$	N/A	$81.9 \pm 11.2\%$	N/A
SVM	$58.9 \pm 7.5\%$	N/A	$44.4 \pm 5.7\%$	N/A
ANN	$82.2 \pm 14.2\%$	N/A	$76.7 \pm 16.8\%$	N/A
Rule Set	$\mathbf{85.5 \pm 13.3\%}$	2.34 ± 0.92	$\mathbf{93.1 \pm 8.1\%}$	2.22 ± 0.58

Table 6-III. mAP for segmentation boxes and masks with different IoU threshold for UM dataset.

IoU	0.5	0.6	0.7	0.8	0.9
box	70.67%	64.41%	49.20%	27.52%	3.24%
mask	69.30%	64.72%	53.07%	33.91%	2.49%

are shown in Table 6-III. The mAP is above 70% when IoU threshold is 50%, which indicates that the segmentation process catches a fairly good number of cancer cells. However, mAP is relatively low with high IoU threshold, because of the low quality of super-pixel-based annotations on the cells' boundary. Figure 6-12 also presents visual results for cell segmentation. The algorithm can easily tell apart cancer cells from blood cells, but the algorithm left some cancer cells with ambiguous boundaries, which are caused by cell overlapping or the wrong focus of the microscopy cameras. We attribute this to the low quality of super-pixels for these cells during annotation. As a result, cells with ambiguous boundaries are usually skipped in annotation if there already exist enough clear cells in the same ROI for cell annotation. The miss detection of cancer cells with ambiguous boundaries is not a big issue for the following cell distribution analysis. There exist numerous cells in each slide. Missing some cells at random will not significantly impact the overall cell composition, and further, the classification performance.

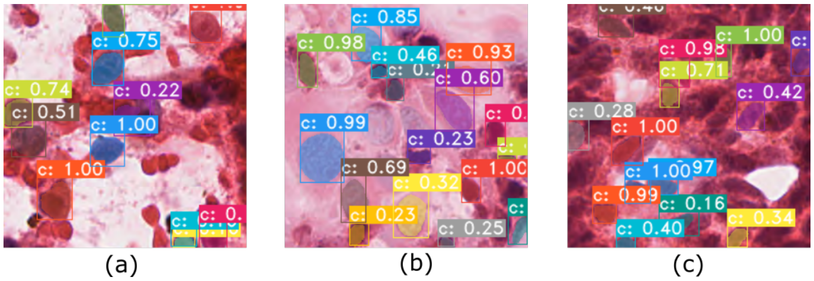


Figure 6-12. Examples of segmentation results. The segmentation network is able to (a) separate cancer cells (purple, large) from blood cells (red, small); (b) segment cells with all sizes, but (c) misses some ambiguous cells. The numbers within the boxes correspond to confidence scores.

6.4 Conclusion

We have presented an automatic but explainable algorithm of cancer subtyping from cytopathology images based on cell composition analysis. This algorithm strictly follows a user-oriented design. The entire pipeline is inspired by the need and knowledge of pathologists from the formative user research. The explainability of the proposed approach is further assessed directly by user studies with pathologists. We included two cancer subtyping tasks, Uveal Melanoma and Cervical Cancer, to evaluate the algorithm's performance. It is worth mentioning that this automatic algorithm is not limited to the tasks mentioned above and is easy to apply for other cancer subtyping tasks with microscopy images. We emphasize the importance of user-oriented designs in cancer subtyping with microscopy images and the proposed algorithm has great potential to offer insights for model designers to build user-oriented explainable models for cancer subtyping and other clinical tasks.

Chapter 7

Conclusion and Future Work

Interpretability is an affordance of interpretable ML systems, i.e., a relationship between models and end users. Therefore, especially in contexts where there exists a high knowledge gap between ML developers and the envisioned end users, developing interpretable ML algorithms without explicitly considering and involving end users may result in products that are unintelligible in the envisioned context and irrelevant in practice. Efforts to build ML systems that afford interpretability in the healthcare context should go beyond computational advances, which is not common practice in the context of interpretable ML for medical image analysis. We acknowledge that building systems that afford interpretability by involving end users in the design process is challenging for medical image analysis and related healthcare tasks. We introduce neural-symbolic reasoning models to achieve the clinical task by implementing clinical guidelines that are commonly agreed on by end users to be useful and understandable. Chapter 3 and Chapter 4 are in the scope of establishing neural symbolic ML models with existing clinical knowledge. We also show how to perform formative user research to iteratively develop clinical evidence to design models for clinical scenarios that end users have no ability to deal with. Furthermore, we present a user study to prove the interpretability and assess the human factors of the designed model, which can be iteratively performed to refine the model design, clinical evidence, and even the entire clinical scenario to be more user friendly and of better human-ML team performance.

Chapter 5 and Chapter 6 fall into addressing the entire procedure to design ML models that afford interpretability in medical image analysis by formative user research, model construction, and empirical user study. This dissertation introduces a new viewpoint to recommend ML designers to actively consider end users during the designing and validating of interpretable machine learning models to make models truly interpretable to end users and achieve higher human-ML performance.

In the detail of my Ph.D. projects, in Chapter 3, we take advantage of a commonly used clinical guideline: hierarchical clinical taxonomy to model label dependencies in Chest X-Rays (CXRs) and present a deep hierarchical multi-label classification approach for CXR CAD. The quantitative results show that the proposed method outperforms the other state-of-the-art approaches, and more importantly, the model outputs strictly follow the hierarchical clinical taxonomy which has great potential to afford interpretability to radiologists. In Chapter 4, we strictly follow the official AAST clinical guidelines to design a splenic injury grading system with Computed Tomography (CT) scans. The most salient features of the grading system, namely active bleeding, pseudoaneurysm, and splenic parenchymal disruption are summarized from the clinical guidelines and are detected and segmented with CNN models. These features are further fed to a rule-based algorithm which is built according to the clinical guidelines to output final grading predictions. Moving forward beyond clinical routine practices, Chapter 5 and Chapter 6 aim to design an interpretable ML algorithm for a clinical task beyond clinical experts' knowledge and ability. The task is predicting UM GEP classes with cytopathology images. Cytopathology images are giga-pixel level images. In order to first determine regions of interest for further analysis, Chapter 5 proposes an automatic but interactive method to efficiently extract high-quality regions in cytopathology images. The method accelerates the region extraction process in cytopathology images more than 10 times compared to manual extraction. The extracted regions with human interaction procedures are more likely to be reliable to pathologists. Moving forward, Chapter 6 uses the regions extracted in Chapter 5 as inputs and established an interpretable algorithm for UM GEP classification. We first perform formative user research to understand pathologists' needs and what they believe is informative to UM prognostication. We found pathologists believe the cell type composition is the most salient feature for UM prognostication. Then, we automatically extract cells and directly analyze cell appearance composition over entire cytopathology images with a rule-based algorithm because pathologists believe cell type distribution already contains adequate information for GEP and further, metastatic risk prediction. Finally, we conduct a user study with pathologists to assess the interpretability of the proposed algorithm.

There also exist multiple future works that are worthy of exploration. First, My Ph.D. projects mainly focus on clinical problems that are relevant to medical images alone. There exist other clinical scenarios that require integrative learning with multiple sources of clinical data, such as medical images, clinical variables, and electronic health records (EHR). Most of these clinical scenarios are closely related to high-stakes decision makings. Thus building interpretable models is highly desired for these clinical scenarios as well. Second, my work aims to afford interpretability to clinical experts, who have a solid training and understanding of the corresponding clinical scenarios. However, other clinical stakeholders, such as the patients and the managers of the hospitals, are also seeking interpretable models to understand the situation of their health and formulate the next steps for the hospitals in the near future. Third, my work established one-time communication between ML models and end users. ML models are the speakers to interpret their own results and end users are the listeners to understand the ML results as a reference to make the final decisions. This is only one approach to the communications between ML models and the end users. An advanced human-machine teaming requires humans and machines to be both receptive listeners and expressive speakers. Effective human-machine collaboration hugely depends on a shared team mental model that includes values, goals, and current states of the task, which has great potential to improve the human-machine teaming experience.

References

1. Vergara, I. A., Norambuena, T., Ferrada, E., Slater, A. W. & Melo, F. StAR: a simple tool for the statistical comparison of ROC curves. *BMC bioinformatics* **9,** 265–265 (June 2008).

2. DeLong, E. R., DeLong, D. M. & Clarke-Pearson, D. L. Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics* **44,** 837–845 (1988).

3. Dekking, F., Kraaikamp, C., Lopuhaä, H. & Meester, L. *A modern introduction to probability and statistics. Understanding why and how* (Springer-Verlag London, 2005).

4. Rennie, J. *ifile: An application of machine learning to e-mail filtering* in *Proc. KDD 2000 workshop on text mining, Boston, MA* (2000).

5. LeCun, Y. *et al.* Learning algorithms for classification: A comparison on handwritten digit recognition. *Neural networks: the statistical mechanics perspective* **261,** 2 (1995).

6. Kubat, M., Holte, R. C. & Matwin, S. Machine learning for the detection of oil spills in satellite radar images. *Machine learning* **30,** 195–215 (1998).

7. Krizhevsky, A., Sutskever, I. & Hinton, G. E. Imagenet classification with deep convolutional neural networks. *Communications of the ACM* **60,** 84–90 (2017).

8. Obermeyer, Z., Powers, B., Vogeli, C. & Mullainathan, S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* **366,** 447–453 (2019).

9. Ghassemi, M., Oakden-Rayner, L. & Beam, A. L. The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health* **3,** e745–e750 (2021).

10. McCoy, L. G., Brenna, C. T., Chen, S. S., Vold, K. & Das, S. Believing in black boxes: Machine learning for healthcare does not need explainability to be evidence-based. *Journal of Clinical Epidemiology* **142,** 252–257 (2022).

11. Vellido, A. The importance of interpretability and visualization in machine learning for applications in medicine and health care. *Neural Computing and Applications* **32,** 18069–18083 (2020).

12. Char, D. S., Abràmoff, M. D. & Feudtner, C. Identifying ethical considerations for machine learning healthcare applications. *The American Journal of Bioethics* **20,** 7–17 (2020).

13. Holzinger, A., Langs, G., Denk, H., Zatloukal, K. & Müller, H. Causability and explainability of artificial intelligence in medicine. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* **9,** e1312 (2019).

14. Salahuddin, Z., Woodruff, H. C., Chatterjee, A. & Lambin, P. Transparency of deep neural networks for medical image analysis: A review of interpretability methods. *Computers in biology and medicine* **140,** 105111 (2022).

15. Norman, D. A. Affordance, conventions, and design. *Interactions* **6,** 38–43 (1999).

16. Chen, H., Gomez, C., Huang, C.-M. & Unberath, M. Explainable medical imaging AI needs human-centered design: guidelines and evidence from a systematic review. *NPJ digital medicine* **5,** 1–15 (2022).

17. Zhang, J., Chen, B., Zhang, L., Ke, X. & Ding, H. Neural, symbolic and neural-symbolic reasoning on knowledge graphs. *AI Open* **2,** 14–35 (2021).

18. Wang, X. *et al. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases* in *Proceedings of the IEEE conference on computer vision and pattern recognition* (2017), 2097–2106.

19. Rajpurkar, P. *et al.* Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225* (2017).

20. Irvin, J. *et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison* in *Proceedings of the AAAI conference on artificial intelligence* **33** (2019), 590–597.

21. Langlotz, C. P. *RadLex: a new method for indexing online educational materials* 2006.

22. Folio, L. R. *Chest imaging: an algorithmic approach to learning* (Springer Science & Business Media, 2012).

23. Demner-Fushman, D., Shooshan, S. E., Rodriguez, L., Antani, S. & Thoma, G. R. *Annotation of chest radiology reports for indexing and retrieval* in *International Workshop on Multimodal Retrieval in the Medical Domain* (2015), 99–111.

24. Chen, H., Miao, S., Xu, D., Hager, G. D. & Harrison, A. P. *Deep Hierarchical Multi-label Classification of Chest X-ray Images* in *Proceedings of The 2nd International Conference on Medical Imaging with Deep Learning* (eds Cardoso, M. J. *et al.*) **102** (PMLR, London, United Kingdom, July 2019), 109–120.

25. Chen, H., Miao, S., Xu, D., Hager, G. D. & Harrison, A. P. Deep hiearchical multi-label classification applied to chest X-ray abnormality taxonomies. *Medical Image Analysis* **66,** 101811 (2020).

26. Dreizin, D. & Munera, F. Blunt polytrauma: evaluation with 64-section whole-body CT angiography. *Radiographics* **32,** 609–631 (2012).

27. Chahine, A. H. *et al.* Management of Splenic Trauma in Contemporary Clinical Practice: A National Trauma Data Bank Study. *Academic Radiology* **28,** S138–S147 (2021).

28. Zarzaur, B. L., Kozar, R. A., Fabian, T. C. & Coimbra, R. A survey of American Association for the Surgery of Trauma member practices in the management of blunt splenic injury. *Journal of Trauma and Acute Care Surgery* **70,** 1026–1031 (2011).

29. Banaste, N. *et al.* Whole-body CT in patients with multiple traumas: factors leading to missed injury. *Radiology* **289,** 374–383 (2018).

30. Watchorn, J., Miles, R. & Moore, N. The role of CT angiography in military trauma. *Clinical radiology* **68,** 39–46 (2013).

31. Glover IV, M., Almeida, R. R., Schaefer, P. W., Lev, M. H. & Mehan Jr, W. A. Quantifying the impact of noninterpretive tasks on radiology report turn-around times. *Journal of the American College of Radiology* **14,** 1498–1503 (2017).

32. Hunter, T. B., Taljanovic, M. S., Krupinski, E., Ovitt, T. & Stubbs, A. Y. Academic radiologists' on-call and late-evening duties. *Journal of the American College of Radiology* **4,** 716–719 (2007).

33. Hanna, T. N., Loehfelm, T., Khosa, F., Rohatgi, S. & Johnson, J.-O. Overnight shift work: factors contributing to diagnostic discrepancies. *Emergency radiology* **23,** 41–47 (2016).

34. Krausz, M. M. & Hirsh, M. Bolus versus continuous fluid resuscitation and splenectomy for treatment of uncontrolled hemorrhagic shock after massive splenic injury. *Journal of Trauma and Acute Care Surgery* **55,** 62–68 (2003).

35. Cai, C. J. *et al. 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* (2019), 1–14.

36. Fails, J. A. & Olsen Jr, D. R. *Interactive machine learning* in *Proceedings of the 8th international conference on Intelligent user interfaces* (2003), 39–45.

37. Kozar, R. A. *et al.* Organ injury scaling 2018 update: spleen, liver, and kidney. *Journal of Trauma and Acute Care Surgery* **85,** 1119–1122 (2018).

38. Chen, H., Unberath, M. & Dreizin, D. Toward automated interpretable AAST grading for blunt splenic injury. *Emergency Radiology,* 1–10 (2022).

39. Singh, A. D., Turell, M. E. & Topham, A. K. Uveal Melanoma: Trends in Incidence, Treatment, and Survival. *Ophthalmology* **118,** 1881–1885 (2011).

40. Grossmann, P. *et al. Defining the biological basis of radiomic phenotypes in lung cancer* in *eLife* (2017).

41. Liu, T. A. *et al.* Gene Expression Profile Prediction in Uveal Melanoma Using Deep Learning: a Pilot Study for Development of an Alternative Survival Prediction Tool. *Ophthalmology Retina* (2020).

42. Chen, H., Liu, T. A., Correa, Z. & Unberath, M. *An Interactive Approach to Region of Interest Selection in Cytologic Analysis of Uveal Melanoma Based on Unsupervised Clustering* in *International Workshop on Ophthalmic Medical Image Analysis* (2020), 114–124.

43. Nourani, M., King, J. & Ragan, E. *The role of domain expertise in user trust and the impact of first impressions with intelligent systems* in *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* **8** (2020), 112–121.

44. Gaube, S. *et al.* Do as AI say: susceptibility in deployment of clinical decision-aids. *npj Digital Medicine* **4,** 1–8 (2021).

45. Chen, H., Liu, T., Gomez, C., Correa, Z. & Unberath, M. An interpretable Algorithm for uveal melanoma subtyping from whole slide cytology images. *arXiv preprint arXiv:2108.06246* (2021).

46. Chen, H. *et al. Anatomy-aware siamese network: Exploiting semantic asymmetry for accurate pelvic fracture detection in x-ray images* in *European Conference on Computer Vision* (2020), 239–255.

47. Liu, T. A., Chen, H., Gomez, C., Correa, Z. M. & Unberath, M. Direct Gene Expression Profile Prediction for Uveal Melanoma from Digital Cytopathology Images via Deep Learning and Salient Image Region Identification. *Ophthalmology Science,* 100240 (2022).

48. Liu, T. Y., Gomez, C., Corrêa, Z. & Unberath, M. Predicting the Gene Expression Profile of Uveal Melanoma Fom Digital Cytopathology via Salient Image Region Identification. *SSRN Electronic Journal* (Jan. 2021).

49. Hubel, D. H. & Wiesel, T. N. Receptive fields and functional architecture of monkey striate cortex. *The Journal of physiology* **195,** 215–243 (1968).

50. Fukushima, K. & Miyake, S. in *Competition and cooperation in neural nets* 267–285 (Springer, 1982).

51. Aerts, H. J. *et al.* Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nature communications* **5,** 1–9 (2014).

52. Lambin, P. *et al.* Radiomics: extracting more information from medical images using advanced feature analysis. *European journal of cancer* **48,** 441–446 (2012).

53. Kortli, Y., Jridi, M., Al Falou, A. & Atri, M. Face recognition systems: A survey. *Sensors* **20,** 342 (2020).

54. Schultheiss, M. *et al.* Lung nodule detection in chest X-rays using synthetic ground-truth data comparing CNN-based diagnosis to human performance. *Scientific Reports* **11,** 1–10 (2021).

55. Huang, G., Liu, Z., Van Der Maaten, L. & Weinberger, K. Q. *Densely connected convolutional networks* in *Proceedings of the IEEE conference on computer vision and pattern recognition* (2017), 4700–4708.

56. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* **1,** 206–215 (2019).

57. Escalante, H. J. *et al. Explainable and interpretable models in computer vision and machine learning* (Springer, 2018).

58. Morch, N. J. *et al. Visualization of neural networks using saliency maps* in *Proceedings of ICNN'95-International Conference on Neural Networks* **4** (1995), 2085–2090.

59. Liu, Z., Gao, J., Yang, G., Zhang, H. & He, Y. Localization and classification of paddy field pests using a saliency map and deep convolutional neural network. *Scientific reports* **6,** 1–12 (2016).

60. Ribeiro, M. T., Singh, S. & Guestrin, C. *" Why should i trust you?" Explaining the predictions of any classifier* in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (2016), 1135–1144.

61. Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions. *Advances in neural information processing systems* **30** (2017).

62. Niu, Z., Zhong, G. & Yu, H. A review on the attention mechanism of deep learning. *Neurocomputing* **452,** 48–62 (2021).

63. Nauta, M., Jutte, A., Provoost, J. & Seifert, C. *This looks like that, because... explaining prototypes for interpretable image recognition* in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* (2021), 441–456.

64. Štrumbelj, E. & Kononenko, I. Explaining prediction models and individual predictions with feature contributions. *Knowledge and information systems* **41,** 647–665 (2014).

65. Dandl, S., Molnar, C., Binder, M. & Bischl, B. *Multi-objective counterfactual explanations* in *International Conference on Parallel Problem Solving from Nature* (2020), 448–469.

66. Mothilal, R. K., Sharma, A. & Tan, C. *Explaining machine learning classifiers through diverse counterfactual explanations* in *Proceedings of the 2020 conference on fairness, accountability, and transparency* (2020), 607–617.

67. Tolomei, G., Silvestri, F., Haines, A. & Lalmas, M. *Interpretable predictions of tree-based ensembles via actionable feature tweaking* in *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining* (2017), 465–474.

68. Lewis, D. *Counterfactuals* (John Wiley & Sons, 2013).

69. Miller, T. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence* **267,** 1–38 (2019).

70. Casalicchio, G., Molnar, C. & Bischl, B. *Visualizing the feature importance for black box models* in *Joint European conference on machine learning and knowledge discovery in databases* (2019), 655–670.

71. Fisher, A., Rudin, C. & Dominici, F. All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously. *J. Mach. Learn. Res.* **20,** 1–81 (2019).

72. Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J. & Wasserman, L. Distribution-free predictive inference for regression. *Journal of the American Statistical Association* **113,** 1094–1111 (2018).

73. Greenwell, B. M., Boehmke, B. C. & McCarthy, A. J. A simple and effective model-based variable importance measure. *arXiv preprint arXiv:1805.04755* (2018).

74. Chen, H., Miao, S., Xu, D., Hager, G. D. & Harrison, A. P. Deep hiearchical multi-label classification applied to chest X-ray abnormality taxonomies. *Medical image analysis* **66,** 101811 (2020).

75. Bortsova, G. *et al. Deep learning from label proportions for emphysema quantification* in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (2018), 768–776.

76. Zimmer, V. A. *et al.* Placenta segmentation in ultrasound imaging: Addressing sources of uncertainty and limited field-of-view. *Medical Image Analysis* **83,** 102639 (2023).

77. Moher, D. *et al.* Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. *Systematic reviews* **4,** 1–9 (2015).

78. Xie, Y., Chen, M., Kao, D., Gao, G. & Chen, X. *CheXplain: Enabling Physicians to Explore and Understand Data-Driven, AI-Enabled Medical Imaging Analysis* in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (2020), 1–13.

79. Jacobs, M. *et al. Designing AI for Trust and Collaboration in Time-Constrained Medical Decisions: A Sociotechnical Lens* in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (2021), 1–14.

80. Molnar, C. *Interpretable machine learning* (Lulu. com, 2020).

81. Abdel Magid, S. *et al.* Channel Embedding for Informative Protein Identification from Highly Multiplexed Images. *23rd International Conference on Medical Image Computing and Computer-Assisted Intervention, MICCAI 2020, October 4, 2020 - October 8, 2020* **12265 LNCS,** 3–13 (2020).

82. Saleem, H., Shahid, A. R. & Raza, B. Visual interpretability in 3D brain tumor segmentation network. English. *Computers in Biology and Medicine* **133** (2021).

83. Shahamat, H. & Saniee Abadeh, M. Brain MRI analysis using a deep learning based evolutionary approach. English. *Neural Networks* **126,** 218–234 (2020).

84. Singla, S. *et al.* Subject2Vec: Generative-Discriminative Approach from a Set of Image Patches to a Vector. *Med Image Comput Comput Assist Interv* **11070,** 502–510 (2018).

85. Sun, J., Darbehani, F., Zaidi, M. & Wang, B. SAUNet: Shape Attentive U-Net for Interpretable Medical Image Segmentation. *23rd International Conference on Medical Image Computing and Computer-Assisted Intervention, MICCAI 2020, October 4, 2020 - October 8, 2020* **12264 LNCS,** 797–806 (2020).

86. Xu, X. *et al.* Automatic glaucoma detection based on transfer induced attention network. English. *Biomedical Engineering Online* **20,** 39 (2021).

87. Yang, H., Kim, J.-Y., Kim, H. & Adhikari, S. P. Guided Soft Attention Network for Classification of Breast Cancer Histopathology Images. *IEEE Transactions on Medical Imaging* **39,** 1306–1315 (2020).

88. Afshar, P. *et al.* MIXCAPS: A capsule network-based mixture of experts for lung nodule malignancy prediction. *Pattern Recognition* **116** (2021).

89. Fan, M., Chakraborti, T., Chang, E. I. C., Xu, Y. & Rittscher, J. Microscopic Fine-Grained Instance Classification Through Deep Attention. *23rd International Conference on Medical Image Computing and Computer-Assisted Intervention, MICCAI 2020, October 4, 2020 - October 8, 2020* **12265 LNCS,** 490–499 (2020).

90. Graziani, M., Lompech, T., Muller, H., Depeursinge, A. & Andrearczyk, V. Interpretable CNN Pruning for Preserving Scale-Covariant Features in Medical Imaging. *3rd International Workshop on Interpretability of Machine Intelligence in Medical Image Computing, iMIMIC 2020, the 2nd International Workshop on Medical Image Learning with Less Labels and Imperfect Data, MIL3ID 2020, and the 5th International Workshop o* **12446 LNCS,** 23–32 (2020).

91. An, F., Li, X. & Ma, X. Medical Image Classification Algorithm Based on Visual Attention Mechanism-MCNN. English. *Oxidative Medicine and Cellular Longevity* **2021** (2021).

92. He, S. *et al.* Multi-channel attention-fusion neural network for brain age estimation: Accuracy, generality, and interpretation with 16,705 healthy MRIs across lifespan. English. *Medical Image Analysis* **72** (2021).

93. Hou, B., Kang, G., Xu, X. & Hu, C. Cross Attention Densely Connected Networks for Multiple Sclerosis Lesion Segmentation. *2019 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2019, November 18, 2019 - November 21, 2019,* 2356–2361 (2019).

94. Huang, Y. & Chung, A. C. S. Evidence localization for pathology images using weakly supervised learning. *22nd International Conference on Medical Image Computing and Computer-Assisted Intervention, MICCAI 2019, October 13, 2019 - October 17, 2019* **11764 LNCS,** 613–621 (2019).

95. Morvan, L. *et al.* Learned Deep Radiomics for Survival Analysis with Attention. *3rd International Workshop on Predictive Intelligence in Medicine, PRIME 2020, held in conjunction with the Medical Image Computing and Computer Assisted Intervention, MICCAI 2020, October 8, 2020 - October 8, 2020* **12329 LNCS,** 35–45 (2020).

96. Human-interpretable image features derived from densely mapped cancer pathology slides predict diverse molecular phenotypes. English. *Nature Communications* **12** (2021).

97. Wongvibulsin, S., Wu, K. C. & Zeger, S. L. Improving Clinical Translation of Machine Learning Approaches Through Clinician-Tailored Visual Displays of Black Box Algorithms: Development and Validation. *JMIR Med Inform* **8,** e15791 (2020).

98. Dong, Y. *et al.* A Polarization-Imaging-Based Machine Learning Framework for Quantitative Pathological Diagnosis of Cervical Precancerous Lesions. English. *IEEE Transactions on Medical Imaging* (2021).

99. Giannini, V., Rosati, S., Regge, D. & Balestra, G. Texture features and artificial neural networks: A way to improve the specificity of a CAD system for multiparametric MR prostate cancer. *14th Mediterranean Conference on Medical and Biological Engineering and Computing, MEDICON 2016, March 31, 2016 - April 2, 2016* **57,** 296–301 (2016).

100. Generate Structured Radiology Report from CT Images Using Image Annotation Techniques: Preliminary Results with Liver CT. *Journal of Digital Imaging* **33,** 375–390 (2020).

101. MacCormick, I. J. C. *et al.* Accurate, fast, data efficient and interpretable glaucoma diagnosis with automated spatial analysis of the whole cup to disc profile. English. *PLoS ONE* **14** (2019).

102. Kunapuli, G. *et al.* A Decision-Support Tool for Renal Mass Classification. English. *Journal of Digital Imaging* **31,** 929–939 (2018).

103. Shen, T., Wang, J., Gou, C. & Wang, F.-Y. Hierarchical Fused Model with Deep Learning and Type-2 Fuzzy Learning for Breast Cancer Diagnosis. *IEEE Transactions on Fuzzy Systems* **28,** 3204–3218 (2020).

104. Li, J., Shi, H. & Hwang, K.-S. An explainable ensemble feedforward method with Gaussian convolutional filter. *Knowledge-Based Systems* **225** (2021).

105. Puyol-Anton, E. *et al.* Assessing the Impact of Blood Pressure on Cardiac Function Using Interpretable Biomarkers and Variational Autoencoders. *10th International Workshop on Statistical Atlases and Computational Models of the Heart, STACOM 2019, held in conjunction with the 22nd International Conference on Medical Image Computing and Computer Assisted Intervention, MICCAI 2019, October 13, 2019* **12009 LNCS,** 22–30 (2020).

106. Lin, Y., Wei, L., Han, S. X., Aberle, D. R. & Hsu, W. EDICNet: An end-to-end detection and interpretable malignancy classification network for pulmonary nodules in computed tomography. *Medical Imaging 2020: Computer-Aided Diagnosis, February 16, 2020 - February 19, 2020* **11314,** The Society of Photo–Optical Instrumentation Engin (2020).

107. Kim, S. T., Lee, H., Kim, H. G. & Ro, Y. M. ICADx: Interpretable computer aided diagnosis of breast masses. *Medical Imaging 2018: Computer-Aided Diagnosis, February 12, 2018 - February 15, 2018* **10575,** DECTRIS Ltd., The Society of Photo–Optical Instrum (2018).

108. Kim, S. T., Lee, J.-H., Lee, H. & Ro, Y. M. Visually interpretable deep network for diagnosis of breast masses on mammograms. English. *Physics in Medicine and Biology* **63,** 235025 (2018).

109. Puyol-Antón, E. *et al.* Interpretable Deep Models for Cardiac Resynchronisation Therapy Response Prediction. *Med Image Comput Comput Assist Interv* **2020,** 284–293 (2020).

110. Wang, C. J. *et al.* Deep learning for liver tumor diagnosis part II: convolutional neural network interpretation using radiologic imaging features. English. *European Radiology* **29,** 3348–3357 (2019).

111. Codella, N. C. F. *et al.* Collaborative human-AI (CHAI): Evidence-based interpretable melanoma classification in dermoscopic images. *1st International Workshop on Machine Learning in Clinical Neuroimaging, MLCN 2018, 1st International Workshop on Deep Learning Fails, DLF 2018, and 1st International Workshop on Interpretability of Machine Intelligence in Medical Image Computing, iMIMIC* **11038 LNCS,** 97–105 (2018).

112. Barata, C., Celebi, M. E. & Marques, J. S. Explainable skin lesion diagnosis using taxonomies. *Pattern Recognition* **110** (2021).

113. Silva, W., Fernandes, K., Cardoso, M. J. & Cardoso, J. S. Towards complementary explanations using deep neural networks. *1st International Workshop on Machine Learning in Clinical Neuroimaging, MLCN 2018, 1st International Workshop on Deep Learning Fails, DLF 2018, and 1st International Workshop on Interpretability of Machine Intelligence in Medical Image Computing, iMIMIC* **11038 LNCS,** 133–140 (2018).

114. Khaleel, M., Tavanapong, W., Wong, J., Oh, J. & De Groen, P. Hierarchical visual concept interpretation for medical image classification. *34th IEEE International Symposium on Computer-Based Medical Systems, CBMS 2021, June 7, 2021 - June 9, 2021* **2021-June,** 25–30 (2021).

115. Pereira, S. *et al.* Enhancing interpretability of automatically extracted machine learning features: application to a RBM-Random Forest system on brain lesion segmentation. English. *Medical Image Analysis* **44,** 228–244 (2018).

116. Yan, K. *et al.* Holistic and comprehensive annotation of clinically significant findings on diverse CT images: Learning from radiology reports and label ontology. *32nd IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2019, June 16, 2019 - June 20, 2019* **2019-June,** 8515–8524 (2019).

117. Verma, A., Shukla, P., Abhishek & Verma, S. An interpretable SVM based model for cancer prediction in mammograms. *1st International Conference on Communication, Networks and Computing, CNC 2018, March 22, 2018 - March 24, 2018* **839,** 443–451 (2019).

118. Li, Y. *et al.* Computer-Aided Cervical Cancer Diagnosis Using Time-Lapsed Colposcopic Images. English. *IEEE Transactions on Medical Imaging* **39,** 3403–3415 (2020).

119. Wang, K. *et al.* A dual-mode deep transfer learning (D2TL) system for breast cancer detection using contrast enhanced digital mammograms. English. *IISE Transactions on Healthcare Systems Engineering* **9,** 357–370 (2019).

120. Zhao, G., Zhou, B., Wang, K., Jiang, R. & Xu, M. Respond-CAM: Analyzing deep models for 3D imaging data by visualizations. *21st International Conference on Medical Image Computing and Computer Assisted Intervention, MICCAI 2018, September 16, 2018 - September 20, 2018* **11070 LNCS,** 485–492 (2018).

121. Folke, T., Yang, S. C.-H., Anderson, S. & Shafto, P. Explainable AI for medical imaging: Explaining pneumothorax diagnoses with Bayesian teaching. *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications III 2021, April 12, 2021 - April 16, 2021* **11746,** The Society of Photo–Optical Instrumentation Engin (2021).

122. Liao, W. *et al.* Clinical Interpretable Deep Learning Model for Glaucoma Diagnosis. English. *IEEE Journal of Biomedical and Health Informatics* **24,** 1405–1412 (2020).

123. Shinde, S., Chougule, T., Saini, J. & Ingalhalikar, M. HR-CAM: Precise localization of pathology using multi-level learning in CNNS. *22nd International Conference on Medical Image Computing and Computer-Assisted Intervention, MICCAI 2019, October 13, 2019 - October 17, 2019* **11767 LNCS,** 298–306 (2019).

124. Ballard, D. H. *Modular learning in neural networks.* in *AAAI* **647** (1987), 279–284.

125. Biffi, C. *et al.* Learning interpretable anatomical features through deep generative models: Application to cardiac remodeling. *21st International Conference on Medical Image Computing and Computer Assisted Intervention, MICCAI 2018, September 16, 2018 - September 20, 2018* **11071 LNCS,** 464–471 (2018).

126. Couteaux, V., Nempont, O., Pizaine, G. & Bloch, I. Towards interpretability of segmentation networks by analyzing deepDreams. *2nd International Workshop on Interpretability of Machine Intelligence in Medical Image Computing, iMIMIC 2019, and the 9th International Workshop on Multimodal Learning for Clinical Decision Support, ML-CDS 2019, held in conjunction with the 22nd Interna* **11797 LNCS,** 56–63 (2019).

127. Guo, X. *et al.* Intelligent medical image grouping through interactive learning. *International Journal of Data Science and Analytics* **2,** 95–105 (2016).

128. Janik, A., Dodd, J., Ifrim, G., Sankaran, K. & Curran, K. Interpretability of a deep learning model in the application of cardiac MRI segmentation with an ACDC challenge dataset. *Medical Imaging 2021: Image Processing, February 15, 2021 - February 19, 2021* **11596,** The Society of Photo–Optical Instrumentation Engin (2021).

129. Sari, C. T. & Gunduz-Demir, C. Unsupervised Feature Extraction via Deep Learning for Histopathological Classification of Colon Tissue Images. *IEEE Transactions on Medical Imaging* **38,** 1139–1149 (2019).

130. Venugopalan, J., Tong, L., Hassanzadeh, H. R. & Wang, M. D. Multimodal deep learning models for early detection of Alzheimer's disease stage. English. *Scientific Reports* **11,** 3254 (2021).

131. Zhu, P. & Ogino, M. Guideline-based additive explanation for computer-aided diagnosis of lung nodules. *2nd International Workshop on Interpretability of Machine Intelligence in Medical Image Computing, iMIMIC 2019, and the 9th International Workshop on Multimodal Learning for Clinical Decision Support, ML-CDS 2019, held in conjunction with the 22nd Interna* **11797 LNCS,** 39–47 (2019).

132. Pirovano, A., Heuberger, H., Berlemont, S., Ladjal, S. & Bloch, I. Improving Interpretability for Computer-Aided Diagnosis Tools on Whole Slide Imaging with Multiple Instance Learning and Gradient-Based Explanations. *3rd International Workshop on Interpretability of Machine Intelligence in Medical Image Computing, iMIMIC 2020, the 2nd International Workshop on Medical Image Learning with Less Labels and Imperfect Data, MIL3ID 2020, and the 5th International Workshop o* **12446 LNCS,** 43–53 (2020).

133. Hao, J., Kosaraju, S. C., Tsaku, N. Z., Song, D. H. & Kang, M. PAGE-Net: Interpretable and Integrative Deep Learning for Survival Analysis Using Histopathological Images and Genomic Data. English. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing* **25,** 355–366 (2020).

134. De Sousa, I. P., Vellasco, M. M. B. R. & da Silva, E. C. Approximate Explanations for Classification of Histopathology Patches. *Workshops of the 20th Joint European Conference on Machine Learning and Knowledge Discovery in Databases, ECML PKDD, September 14, 2020 - September 18, 2020* **1323,** 517–526 (2020).

135. Li, X., Dvornek, N. C., Zhuang, J., Ventola, P. & Duncan, J. S. Brain biomarker interpretation in ASD using deep learning and fMRI. *21st International Conference on Medical Image Computing and Computer Assisted Intervention, MICCAI 2018, September 16, 2018 - September 20, 2018* **11072 LNCS,** 206–214 (2018).

136. Quellec, G. *et al.* ExplAIn: Explanatory artificial intelligence for diabetic retinopathy diagnosis. English. *Medical Image Analysis* **72** (2021).

137. Uzunova, H., Ehrhardt, J., Kepp, T. & Handels, H. Interpretable explanations of black box classifiers applied on medical images by meaningful perturbations using variational autoencoders. *Medical Imaging 2019: Image Processing, February 19, 2019 - February 21, 2019* **10949,** The Society of Photo–Optical Instrumentation Engin (2019).

138. Liu, J. *et al.* Ultrasound Liver Fibrosis Diagnosis Using Multi-indicator Guided Deep Neural Networks. *10th International Workshop on Machine Learning in Medical Imaging, MLMI 2019 held in conjunction with the 22nd International Conference on Medical Image Computing and Computer-Assisted Intervention, MICCAI 2019, October 13, 2019 - October 13, 2019* **11861 LNCS,** 230–237 (2019).

139. Liu, Y. *et al.* Act Like a Radiologist: Towards Reliable Multi-view Correspondence Reasoning for Mammogram Mass Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021).

140. Oktay, O. *et al.* Anatomically Constrained Neural Networks (ACNNs): Application to Cardiac Image Enhancement and Segmentation. *IEEE Transactions on Medical Imaging* **37,** 384–395 (2018).

141. Peng, T., Boxberg, M., Weichert, W., Navab, N. & Marr, C. Multi-task learning of a deep K-nearest neighbour network for histopathological image classification and retrieval. *22nd International Conference on Medical Image Computing and Computer-Assisted Intervention, MICCAI 2019, October 13, 2019 - October 17, 2019* **11764 LNCS,** 676–684 (2019).

142. Liu, Y., Li, Z., Ge, Q., Lin, N. & Xiong, M. Deep Feature Selection and Causal Analysis of Alzheimer's Disease. English. *Frontiers in Neuroscience* **13** (2019).

143. Ren, H. *et al.* Interpretable Pneumonia Detection by Combining Deep Learning and Explainable Models with Multisource Data. *Ieee Access* **9,** 95872–95883 (2021).

144. Velikova, M., Lucas, P. J. F., Samulski, M. & Karssemeijer, N. On the interplay of machine learning and background knowledge in image interpretation by Bayesian networks. English. *Artificial Intelligence in Medicine* **57,** 73–86 (2013).

145. Carneiro, G., Zorron Cheng Tao Pu, L., Singh, R. & Burt, A. Deep learning uncertainty and confidence calibration for the five-class polyp classification from colonoscopy. English. *Medical Image Analysis* **62** (2020).

146. Sabol, P. *et al.* Explainable classifier for improving the accountability in decision-making for colorectal cancer diagnosis from histopathological images. English. *Journal of Biomedical Informatics* **109** (2020).

147. Tanno, R. *et al.* Uncertainty modelling in deep learning for safer neuroimage enhancement: Demonstration in diffusion MRI. English. *NeuroImage* **225** (2021).

148. Doshi-Velez, F. & Kim, B. Towards a rigorous science of interpretable machine learning. *preprint at https://arxiv.org/abs/1702.08608* (2017).

149. Adebayo, J. *et al. Sanity Checks for Saliency Maps* in *Advances in Neural Information Processing Systems* (eds Bengio, S. *et al.*) **31** (Curran Associates, Inc., 2018).

150. Yeche, H., Harrison, J. & Berthier, T. UBS: A dimension-agnostic metric for concept vector interpretability applied to radiomics. *2nd International Workshop on Interpretability of Machine Intelligence in Medical Image Computing, iMIMIC 2019, and the 9th International Workshop on Multimodal Learning for Clinical Decision Support, ML-CDS 2019, held in conjunction with the 22nd Interna* **11797 LNCS,** 12–20 (2019).

151. Chen, H., Miao, S., Xu, D., Hager, G. D. & Harrison, A. P. *Deep hierarchical multi-label classification of chest X-ray images* in *International conference on medical imaging with deep learning* (2019), 109–120.

152. Mettler, F. A. *et al.* Radiologic and Nuclear Medicine Studies in the United States and Worldwide: Frequency, Radiation Dose, and Comparison with Other Radiation Sources—1950–2007. *Radiology* **253.** PMID: 19789227, 520–531. eprint: https://doi.org/10.1148/radiol.2532082010 (2009).

153. Jaeger, S. *et al.* Automatic screening for tuberculosis in chest radiographs: a survey. eng. *Quantitative Imaging in Medicine and Surgery* **3,** 89–99 (Apr. 2013).

154. Wang, X. *et al. ChestX-Ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases* in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (July 2017), 3462–3471.

155. Yao, L. *et al.* Learning to diagnose from scratch by exploiting dependencies among labels. *CoRR* **abs/1710.10501.** arXiv: 1710.10501 (2017).

156. Gündel, S. *et al.* Multi-task Learning for Chest X-ray Abnormality Classification on Noisy Labels. *arXiv:1905.06362 [cs].* arXiv: 1905.06362 (May 2019).

157. Irvin, J. *et al.* CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison. *Proceedings of the AAAI Conference on Artificial Intelligence* **33,** 590–597 (July 2019).

158. Bustos, A., Pertusa, A., Salinas, J. M. & de la Iglesia-Vayá, M. PadChest: A large chest x-ray image dataset with multi-label annotated reports. *ArXiv* **abs/1901.07441** (2019).

159. Johnson, A. E. W. *et al.* MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. en. *Scientific Data* **6,** 1–8 (Dec. 2019).

160. Islam, M. T., Aowal, M. A., Minhaz, A. T. & Ashraf, K. Abnormality Detection and Localization in Chest X-Rays using Deep Convolutional Neural Networks. *arXiv:1705.09850 [cs].* arXiv: 1705.09850 (May 2017).

161. Guan, Q. *et al.* Diagnose like a Radiologist: Attention Guided Convolutional Neural Network for Thorax Disease Classification. *arXiv:1801.09927 [cs].* arXiv: 1801.09927 (Jan. 2018).

162. Wang, H. & Xia, Y. ChestNet: A Deep Neural Network for Classification of Thoracic Diseases on Chest Radiography. *arXiv:1807.03058 [cs].* arXiv: 1807.03058 (July 2018).

163. Liu, H. *et al.* SDFN: Segmentation-based deep fusion network for thoracic disease classification in chest X-ray images. *Computerized Medical Imaging and Graphics* **75,** 66–73 (2019).

164. Yan, C., Yao, J., Li, R., Xu, Z. & Huang, J. *Weakly Supervised Deep Learning for Thoracic Disease Classification and Localization on Chest X-rays* in *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics* event-place: Washington, DC, USA (ACM, New York, NY, USA, 2018), 103–110.

165. Li, Z. *et al. Thoracic Disease Identification and Localization with Limited Supervision* en. in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (IEEE, Salt Lake City, UT, June 2018), 8290–8299.

166. Cai, J. *et al. Iterative Attention Mining for Weakly Supervised Thoracic Disease Pattern Localization in Chest X-Rays* in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018* (eds Frangi, A. F., Schnabel, J. A., Davatzikos, C., Alberola-López, C. & Fichtinger, G.) (Springer International Publishing, Cham, 2018), 589–598.

167. Zhang, M.-L. & Zhou, Z.-H. A Review on Multi-Label Learning Algorithms. en. *IEEE Transactions on Knowledge and Data Engineering* **26,** 1819–1837 (Aug. 2014).

168. Dembczyński, K., Waegeman, W., Cheng, W. & Hüllermeier, E. On Label Dependence and Loss Minimization in Multi-label Classification. *Mach. Learn.* **88,** 5–45 (July 2012).

169. Stevens, R. *et al.* Using OWL to model biological knowledge. *International Journal of Human-Computer Studies* **65,** 583–594 (2007).

170. Humphreys, B. L. & Lindberg, D. A. The UMLS project: making the conceptual connection between users and the information they need. *Bulletin of the Medical Library Association* **81,** 170–177 (Apr. 1993).

171. Stearns, M. Q., Price, C., Spackman, K. A. & Wang, A. Y. SNOMED clinical terms: overview of the development process and project status. *Proceedings of the AMIA Symposium,* 662–666 (2001).

172. Langlotz, C. P. RadLex: A New Method for Indexing Online Educational Materials. *RadioGraphics* **26.** PMID: 17102038, 1595–1597 (2006).

173. Folio, L. *Chest imaging: An algorithmic approach to learning* 1–147 (Springer, Jan. 2012).

174. Demner-Fushman, D., Shooshan, S. E., Rodriguez, L., Antani, S. & Thoma, G. R. *Annotation of Chest Radiology Reports for Indexing and Retrieval* in *Multimodal Retrieval in the Medical Domain* (eds Müller, H., Jimenez del Toro, O. A., Hanbury, A., Langs, G. & Foncubierta Rodriguez, A.) (Springer International Publishing, Cham, 2015), 99–111.

175. Dimitrovski, I., Kocev, D., Loskovska, S. & Deroski, S. Hierarchical Annotation of Medical Images. *Pattern Recogn.* **44,** 2436–2449 (Oct. 2011).

176. Bi, W. & Kwok, J. T. Bayes-Optimal Hierarchical Multilabel Classification. *IEEE Transactions on Knowledge and Data Engineering* **27,** 2907–2918 (Nov. 2015).

177. McCallum, A., Rosenfeld, R., Mitchell, T. M. & Ng, A. Y. *Improving Text Classification by Shrinkage in a Hierarchy of Classes* in *ICML* (1998).

178. Cesa-bianchi, N., Gentile, C., Tironi, A. & Zaniboni, L. Incremental Algorithms for Hierarchical Classification (eds Saul, L. K., Weiss, Y. & Bottou, L.) 233–240 (2005).

179. Cai, L. Exploiting Known Taxonomies in Learning Overlapping Concepts. en, 6 (2007).

180. Vens, C., Struyf, J., Schietgat, L., Džeroski, S. & Blockeel, H. Decision trees for hierarchical multi-label classification. en. *Machine Learning* **73,** 185 (Aug. 2008).

181. Redmon, J. & Farhadi, A. *YOLO9000: Better, Faster, Stronger* in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* **00** (July 2017), 6517–6525.

182. Roy, D., Panda, P. & Roy, K. Tree-CNN: A hierarchical Deep Convolutional Neural Network for incremental learning. *Neural Networks* **121,** 148–160 (Jan. 2020).

183. Yan, Z. *et al. HD-CNN: Hierarchical Deep Convolutional Neural Networks for Large Scale Visual Recognition* in *2015 IEEE International Conference on Computer Vision (ICCV)* (Dec. 2015), 2740–2748.

184. Guo, Y., Liu, Y., Bakker, E. M., Guo, Y. & Lew, M. S. CNN-RNN: a large-scale hierarchical image classification framework. en. *Multimedia Tools and Applications* **77,** 10251–10271 (Apr. 2018).

185. Kowsari, K. *et al. HDLTex: Hierarchical Deep Learning for Text Classification* in *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)* (Dec. 2017), 364–371.

186. Pourghassem, H. & Ghassemian, H. Content-based medical image classification using a new hierarchical merging scheme. *Computerized Medical Imaging and Graphics* **32,** 651–661 (2008).

187. Kohli, M. D., Summers, R. M. & Geis, J. R. *Medical Image Data and Datasets in the Era of Machine Learning: Whitepaper from the 2016 C-MIMI Meeting Dataset Session* in *Journal of Digital Imaging* (2017).

188. Harvey, H. & Glocker, B. in *Artificial Intelligence in Medical Imaging* 61–72 (Springer, 2019).

189. Erdi, C., Ecem, S., Ernst, T. S., Keelin, M. & Bram, v. G. *Handling label noise through model confidence and uncertainty: application to chest radiograph classification* 2019.

190. Yu, H.-F., Jain, P., Kar, P. & Dhillon, I. *Large-scale Multi-label Learning with Missing Labels* in *Proceedings of the 31st International Conference on Machine Learning* (eds Xing, E. P. & Jebara, T.) **32** (PMLR, Bejing, China, June 2014), 593–601.

191. Kong, X. *et al.* in *Proceedings of the 2014 SIAM International Conference on Data Mining* 920–928 (2014).

192. Zhao, F. & Guo, Y. *Semi-supervised Multi-label Learning with Incomplete Labels* in *Proceedings of the 24th International Conference on Artificial Intelligence* (AAAI Press, Buenos Aires, Argentina, 2015), 4062–4068.

193. Elkan, C. & Noto, K. *Learning classifiers from only positive and unlabeled data* en. in *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD 08* (ACM Press, Las Vegas, Nevada, USA, 2008), 213.

194. Liu, B., Dai, Y., Li, X., Lee, W. S. & Yu, P. S. *Building text classifiers using positive and unlabeled examples* in *Third IEEE International Conference on Data Mining* (Nov. 2003), 179–186.

195. Qi, Z., Yang, M. W., Zhang, Z. & Zhang, Z. *Mining partially annotated images* in *KDD* (2011), 1199–1207.

196. Bucak, S. S., Jin, R. & Jain, A. K. *Multi-label learning with incomplete class assignments* en. in *CVPR 2011* (IEEE, Colorado Springs, CO, USA, June 2011), 2801–2808.

197. Yang, S.-J., Jiang, Y. & Zhou, Z.-H. Multi-Instance Multi-Label Learning with Weak Label. en, 7 (2013).

198. Gohagan, J. K., Prorok, P. C., Hayes, R. B. & Kramer, B.-S. The Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial of the National Cancer Institute: History, organization, and status. *Controlled Clinical Trials* **21,** 251S–272S (2000).

199. Peng, Y. *et al. NegBio: a high-performance tool for negation and uncertainty detection in radiology reports* in *AMIA 2018 Informatics Summit 2018* (2018).

200. Oakden-Rayner, L. Exploring large scale public medical image datasets. *Academic radiology* (2019).

201. Huang, G., Liu, Z., v. d. Maaten, L. & Weinberger, K. Q. *Densely Connected Convolutional Networks* in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (July 2017), 2261–2269.

202. Deng, J. *et al. ImageNet: A Large-Scale Hierarchical Image Database* in *CVPR09* (2009).

203. Gündel, S. *et al. Learning to Recognize Abnormalities in Chest X-Rays with Location-Aware Dense Networks* in *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications* (eds Vera-Rodriguez, R., Fierrez, J. & Morales, A.) (Springer International Publishing, Cham, Jan. 2019), 757–765.

204. Dumais, S. & Chen, H. *Hierarchical classification of Web content* en. in *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '00* (ACM Press, Athens, Greece, 2000), 256–263.

205. Dreizin, D. *et al.* Blunt splenic injury: Assessment of follow-up CT utility using quantitative volumetry. *Frontiers in radiology,* 23 (2022).

206. Zarzaur, B. L. *et al.* The splenic injury outcomes trial: an American Association for the Surgery of Trauma multi-institutional study. *Journal of Trauma and Acute Care Surgery* **79,** 335–342 (2015).

207. Haan, J. M. *et al.* Splenic embolization revisited: a multicenter review. *Journal of Trauma and Acute Care Surgery* **56,** 542–547 (2004).

208. Barquist, E. S. *et al.* Inter-and intrarater reliability in computed axial tomographic grading of splenic injury: why so many grading scales? *Journal of Trauma and Acute Care Surgery* **56,** 334–338 (2004).

209. Clark, R., Hird, K., Misur, P., Ramsay, D. & Mendelson, R. CT grading scales for splenic injury: Why can't we agree? *Journal of Medical Imaging and Radiation Oncology* **55,** 163–169 (2011).

210. Cruz-Romero, C., Agarwal, S., Abujudeh, H. H., Thrall, J. & Hahn, P. F. Spleen volume on CT and the effect of abdominal trauma. *Emergency radiology* **23,** 315–323 (2016).

211. Wood, A. *et al. Fully automated spleen localization and segmentation using machine learning and 3D active contours* in *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (2018), 53–56.

212. Dandin, O. *et al.* Automated segmentation of the injured spleen. *International journal of computer assisted radiology and surgery* **11,** 351–368 (2016).

213. Wang, J., Wood, A., Gao, C., Najarian, K. & Gryak, J. Automated Spleen Injury Detection Using 3D Active Contours and Machine Learning. *Entropy* **23,** 382 (2021).

214. Teomete, U. *et al.* Automated computer-aided diagnosis of splenic lesions due to abdominal trauma. *Hippokratia* **22,** 80 (2018).

215. DeGrave, A. J., Janizek, J. D. & Lee, S.-I. AI for radiographic COVID-19 detection selects shortcuts over signal. *Nature Machine Intelligence* **3,** 610–619 (2021).

216. Zapaishchykova, A. *et al. An interpretable approach to automated severity scoring in pelvic trauma* in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (2021), 424–433.

217. Vlontzos, A., Rueckert, D. & Kainz, B. A review of causality for learning algorithms in medical image analysis. *arXiv preprint arXiv:2206.05498* (2022).

218. Arrieta, A. B. *et al.* Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information fusion* **58,** 82–115 (2020).

219. Boscak, A. R. *et al.* Optimizing trauma multidetector CT protocol for blunt splenic injury: need for arterial and portal venous phase scans. *Radiology* **268,** 79–88 (2013).

220. Uyeda, J. W., LeBedis, C. A., Penn, D. R., Soto, J. A. & Anderson, S. W. Active hemorrhage and vascular injuries in splenic trauma: utility of the arterial phase in multidetector CT. *Radiology* **270,** 99–106 (2014).

221. Zhou, Y. *et al.* External Attention Assisted Multi-Phase Splenic Vascular Injury Segmentation With Limited Data. *IEEE Transactions on Medical Imaging* **41,** 1346–1357 (2021).

222. Antonelli, M. *et al.* The medical segmentation decathlon. *Nature communications* **13,** 1–13 (2022).

223. Cooney, R. *et al.* Limitations of splenic angioembolization in treating blunt splenic injury. *Journal of Trauma and Acute Care Surgery* **59,** 926–932 (2005).

224. Bhullar, I. S. *et al.* Selective angiographic embolization of blunt splenic traumatic injuries in adults decreases failure rate of nonoperative management. *Journal of Trauma and Acute Care Surgery* **72,** 1127–1134 (2012).

225. Crichton, J. C. I., Naidoo, K., Yet, B., Brundage, S. I. & Perkins, Z. The role of splenic angioembolization as an adjunct to nonoperative management of blunt splenic injuries: a systematic review and meta-analysis. *Journal of Trauma and Acute Care Surgery* **83,** 934–943 (2017).

226. Requarth, J. A., D'Agostino Jr, R. B. & Miller, P. R. Nonoperative management of adult blunt splenic injury with and without splenic artery embolotherapy: a meta-analysis. *Journal of Trauma and Acute Care Surgery* **71,** 898–903 (2011).

227. Moore, E. E. *et al.* Organ injury scaling: spleen and liver (1994 revision). *Journal of Trauma and Acute Care Surgery* **38,** 323–324 (1995).

228. Haan, J. M., Bochicchio, G. V., Kramer, N. & Scalea, T. M. Nonoperative management of blunt splenic injury: a 5-year experience. *Journal of Trauma and Acute Care Surgery* **58,** 492–498 (2005).

229. Miller, P. R. *et al.* Prospective trial of angiography and embolization for all grade III to V blunt splenic injuries: nonoperative management success rate is significantly improved. *Journal of the American College of Surgeons* **218,** 644–648 (2014).

230. Ren, S., He, K., Girshick, R. & Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems* **28** (2015).

231. Isensee, F., Jaeger, P. F., Kohl, S. A., Petersen, J. & Maier-Hein, K. H. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods* **18,** 203–211 (2021).

232. Xie, Q., Luong, M.-T., Hovy, E. & Le, Q. V. *Self-training with noisy student improves imagenet classification* in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2020), 10687–10698.

233. Consortium, M. *MONAI: Medical Open Network for AI* version 1.0.1. If you use this software, please cite it using these metadata. Oct. 2022.

234. Ronneberger, O., Fischer, P. & Brox, T. *U-net: Convolutional networks for biomedical image segmentation* in *International Conference on Medical image computing and computer-assisted intervention* (2015), 234–241.

235. Lin, T.-Y. *et al. Microsoft COCO: Common Objects in Context* cite arxiv:1405.0312Comment: 1) updated annotation pipeline description and figures; 2) added new section describing datasets splits; 3) updated author list. 2014.

236. Zhang, R., Tian, Z., Shen, C., You, M. & Yan, Y. *Mask encoding for single shot instance segmentation* in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2020), 10226–10235.

237. Wu, Y., Kirillov, A., Massa, F., Lo, W.-Y. & Girshick, R. *Detectron2* `https://github.com/facebookresearch/detectron2`. 2019.

238. Bhangu, A., Nepogodiev, D., Lal, N. & Bowley, D. M. Meta-analysis of predictive factors and outcomes for failure of non-operative management of blunt splenic trauma. *Injury* **43,** 1337–1346 (2012).

239. Dreizin, D. *et al.* Added value of deep learning-based liver parenchymal CT volumetry for predicting major arterial injury after blunt hepatic trauma: a decision tree analysis. *Abdominal Radiology* **46,** 2556–2566 (2021).

240. Dreizin, D. *et al.* A multiscale deep learning method for quantitative visualization of traumatic hemoperitoneum at CT: assessment of feasibility and comparison with subjective categorical estimation. *Radiology: Artificial Intelligence* **2** (2020).

241. Dreizin, D. *et al.* Deep learning-based quantitative visualization and measurement of extraperitoneal hematoma volumes in patients with pelvic fractures: potential role in personalized forecasting and decision support. *The journal of trauma and acute care surgery* **88,** 425 (2020).

242. Dreizin, D., Zhou, Y., Zhang, Y., Tirada, N. & Yuille, A. L. Performance of a deep learning algorithm for automated segmentation and quantification of traumatic pelvic hematomas on CT. *Journal of digital imaging* **33,** 243–251 (2020).

243. Dreizin, D. *et al.* A pilot study of deep learning-based CT volumetry for traumatic hemothorax. *Emergency Radiology,* 1–8 (2022).

244. Landis, J. R. & Koch, G. G. The measurement of observer agreement for categorical data. *biometrics,* 159–174 (1977).

245. Morell-Hofert, D. *et al.* Validation of the revised 2018 AAST-OIS classification and the CT severity index for prediction of operative management and survival in patients with blunt spleen and liver injuries. *European Radiology* **30,** 6570–6581 (2020).

246. Jeavons, C., Hacking, C., Beenen, L. F. & Gunn, M. L. A review of split-bolus single-pass CT in the assessment of trauma patients. *Emergency radiology* **25,** 367–374 (2018).

247. Beenen, L. F. *et al.* Split bolus technique in polytrauma: a prospective study on scan protocols for trauma analysis. *Acta radiologica* **56,** 873–880 (2015).

248. Lopez Jr, J. M. *et al.* Subcapsular hematoma in blunt splenic injury: A significant predictor of failure of nonoperative management. *Journal of Trauma and Acute Care Surgery* **79,** 957–960 (2015).

249. Scatamacchia, S. A., Raptopoulos, V., Fink, M. P. & Silva, W. E. Splenic trauma in adults: impact of CT grading on management. *Radiology* **171,** 725–729 (1989).

250. Rowell, S. E. *et al.* Western Trauma Association Critical Decisions in Trauma: Management of adult blunt splenic trauma—2016 updates. *Journal of Trauma and Acute Care Surgery* **82,** 787–793 (2017).

251. Coccolini, F. *et al.* Splenic trauma: WSES classification and guidelines for adult and pediatric patients. *World Journal of Emergency Surgery* **12,** 1–26 (2017).

252. Boscak, A. & Shanmuganathan, K. Splenic trauma: what is new? *Radiologic Clinics* **50,** 105–122 (2012).

253. Lee, J. T. *et al.* American Society of Emergency Radiology multicenter blunt splenic trauma study: CT and clinical findings. *Radiology* **299,** 122–130 (2021).

254. Corrêa, Z. & Augsburger, J. Sufficiency of FNAB aspirates of posterior uveal melanoma for cytologic versus GEP classification in 159 patients, and relative prognostic significance of these classifications. *Graefe's archive for clinical and experimental ophthalmology = Albrecht von Graefes Archiv fur klinische und experimentelle Ophthalmologie* **252** (Nov. 2013).

255. Folberg, R., Augsburger, J. J., Gamel, J. W., Shields, J. A. & Lang, W. R. Fine-Needle Aspirates of Uveal Melanomas and Prognosis. *American Journal of Ophthalmology* **100,** 654–657 (1985).

256. Roullier, V., Lézoray, O., Ta, V.-T. & Elmoataz, A. Multi-resolution graph-based analysis of histopathological whole slide images: Application to mitotic cell extraction and visualization. *Computerized Medical Imaging and Graphics* **35.** Whole Slide Image Process, 603–615 (2011).

257. Barker, J., Hoogi, A., Depeursinge, A. & Rubin, D. L. Automated classification of brain tumor type in whole-slide digital pathology images using local representative tiles. *Medical image analysis* **30,** 60–71 (2016).

258. Lin, H. *et al.* *ScanNet: A Fast and Dense Scanning Framework for Metastastic Breast Cancer Detection from Whole-Slide Image* in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)* (Mar. 2018), 539–546.

259. Li, J. *et al.* An attention-based multi-resolution model for prostate whole slide imageclassification and localization. *CoRR* **abs/1905.13208.** arXiv: 1905.13208 (2019).

260. Dov, D. *et al.* *A Deep-Learning Algorithm for Thyroid Malignancy Prediction From Whole Slide Cytopathology Images* 2019. arXiv: 1904.12739 [physics.med-ph].

261. Garud, H. *et al.* *High-Magnification Multi-views Based Classification of Breast Fine Needle Aspiration Cytology Cell Samples Using Fusion of Decisions from Deep Convolutional Networks* in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (July 2017), 828–833.

262. Saikia, A. R., Bora, K., Mahanta, L. B. & Das, A. K. Comparative assessment of CNN architectures for classification of breast FNAC images. *Tissue and Cell* **57.** EM in cell and tissues, 8–14 (2019).

263. Zhu, Z., Xia, Y., Shen, W., Fishman, E. & Yuille, A. *A 3D Coarse-to-Fine Framework for Volumetric Medical Image Segmentation* in *2018 International Conference on 3D Vision (3DV)* (Sept. 2018), 682–690.

264. Zhu, Z., Xia, Y., Xie, L., Fishman, E. K. & Yuille, A. L. Multi-Scale Coarse-to-Fine Segmentation for Screening Pancreatic Ductal Adenocarcinoma. *CoRR* **abs/1807.02941.** arXiv: 1807.02941 (2018).

265. Zhou, Y. *et al.* in *Deep Learning and Convolutional Neural Networks for Medical Imaging and Clinical Informatics* 43–67 (Springer International Publishing, Cham, 2019).

266. Chang, L., Zhang, M. & Li, W. *A coarse-to-fine approach for medical hyperspectral image classification with sparse representation* in *AOPC 2017: Optical Spectroscopy and Imaging* (eds Yu, J. *et al.*) **10461** (SPIE, 2017), 136–144.

267. Liu, J., Chen, F., Shi, H. & Liao, H. *Single Image Super-Resolution for MRI Using a Coarse-to-Fine Network* in *2nd International Conference for Innovation in Biomedical Engineering and Life Sciences* (eds Ibrahim, F., Usman, J., Ahmad, M. Y., Hamzah, N. & Teh, S. J.) (Springer Singapore, Singapore, 2018), 241–245.

268. Wang, G. *et al.* Interactive Medical Image Segmentation Using Deep Learning With Image-Specific Fine Tuning. *IEEE Transactions on Medical Imaging* **37,** 1562–1573 (July 2018).

269. Xu, N., Price, B., Cohen, S., Yang, J. & Huang, T. S. *Deep Interactive Object Selection* in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2016).

270. Girard, N., Zhygallo, A. & Tarabalka, Y. *ClusterNet: unsupervised generic feature learning for fast interactive satellite image segmentation* in *Image and Signal Processing for Remote Sensing XXV* (eds Bruzzone, L. & Bovolo, F.) **11155** (SPIE, 2019), 244–254.

271. Aresta, G. *et al.* iW-Net: an automatic and minimalistic interactive lung nodule segmentation deep network. *Scientific Reports* **9** (Dec. 2019).

272. Amrehn, M. *et al.* *UI-Net: Interactive Artificial Neural Networks for Iterative Image Segmentation Based on a User Model* in (Sept. 2017).

273. Brendel, W. & Bethge, M. *Approximating CNNs with Bag-of-local-Features models works surprisingly well on ImageNet* in *International Conference on Learning Representations* (2019).

274. Yang, B., Fu, X., Sidiropoulos, N. & Hong, M. *Towards K-means-friendly spaces: Simultaneous deep learning and clustering* English (US). in *34th International Conference on Machine Learning, ICML 2017* (International Machine Learning Society (IMLS), Jan. 2017), 5888–5901.

275. Caron, M., Bojanowski, P., Joulin, A. & Douze, M. *Deep Clustering for Unsupervised Learning of Visual Features* in *Computer Vision – ECCV 2018* (eds Ferrari, V., Hebert, M., Sminchisescu, C. & Weiss, Y.) (Springer International Publishing, Cham, 2018), 139–156.

276. Paszke, A. *et al.* Automatic differentiation in PyTorch (2017).

277. Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

278. Breugom, A. J. *et al.* Adjuvant chemotherapy and relative survival of patients with stage II colon cancer–a EURECCA international comparison between the Netherlands, Denmark, Sweden, England, Ireland, Belgium, and Lithuania. *European journal of cancer* **63,** 110–117 (2016).

279. Dotan, E. & Cohen, S. J. *Challenges in the management of stage II colon cancer* in *Seminars in oncology* **38** (2011), 511–520.

280. Brancati, N. *et al.* Bracs: A dataset for breast carcinoma subtyping in h&e histology images. *Database* **2022** (2022).

281. Liu, Y. *et al.* Detecting cancer metastases on gigapixel pathology images. *arXiv preprint arXiv:1703.02442* (2017).

282. Hou, L. *et al.* *Patch-based convolutional neural network for whole slide tissue image classification* in *Proceedings of the IEEE conference on computer vision and pattern recognition* (2016), 2424–2433.

283. Xu, Y. *et al.* Large scale tissue histopathology image classification, segmentation, and visualization via deep convolutional activation features. *BMC bioinformatics* **18,** 1–17 (2017).

284. Zhang, H. *et al.* Piloting a deep learning model for predicting nuclear BAP1 immunohistochemical expression of uveal melanoma from hematoxylin-and-eosin sections. *Translational Vision Science & Technology* **9,** 50–50 (2020).

285. Chikontwe, P., Kim, M., Nam, S. J., Go, H. & Park, S. H. *Multiple Instance Learning with Center Embeddings for Histopathology Classification* in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (2020), 519–528.

286. Campanella, G. *et al.* Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature medicine* **25,** 1301–1309 (2019).

287. Onken, M. D., Worley, L. A., Ehlers, J. P. & Harbour, J. W. Gene expression profiling in uveal melanoma reveals two molecular classes and predicts metastatic death. *Cancer research* **64,** 7205–7209 (2004).

288. Guo, Q. *et al.* MEF2C-AS1 regulates its nearby gene MEF2C to mediate cervical cancer cell malignant phenotypes in vitro. *Biochemical and Biophysical Research Communications* **632,** 48–54 (2022).

289. Verhoef, V. M. *et al.* Methylation marker analysis and HPV16/18 genotyping in high-risk HPV positive self-sampled specimens to identify women with high grade CIN or cervical cancer. *Gynecologic oncology* **135,** 58–63 (2014).

290. Achanta, R. *et al.* *SLIC Superpixels* 2010.

291. Bolya, D., Zhou, C., Xiao, F. & Lee, Y. J. *YOLACT: Real-Time Instance Segmentation* in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (Oct. 2019).

292. McInnes, L., Healy, J. & Melville, J. *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction* 2020. arXiv: `1802.03426 [stat.ML]`.

293. Pearson, K. X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* **50,** 157–175 (1900).

294. Kolmogorov, A. Sulla determinazione empirica di una lgge di distribuzione. *Inst. Ital. Attuari, Giorn.* **4,** 83–91 (1933).

295. Wang, T. *et al.* A bayesian framework for learning rule sets for interpretable classification. *The Journal of Machine Learning Research* **18,** 2357–2393 (2017).

296. Hussain, E., Mahanta, L. B., Borah, H. & Das, C. R. Liquid based-cytology pap smear dataset for automated multi-class diagnosis of pre-cancerous and cervical cancer lesions. *Data in brief* **30,** 105589 (2020).

297. Kim, A. FastSLIC: Optimized SLIC Superpixel.

298. Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

299. Grinberg, M. *Flask web development: developing web applications with python* (" O'Reilly Media, Inc.", 2018).

300. Coudray, N. *et al.* Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. *Nature medicine* **24,** 1559–1567 (2018).

301. He, K., Zhang, X., Ren, S. & Sun, J. Deep Residual Learning for Image Recognition. *arXiv preprint arXiv:1512.03385* (2015).

Curriculum Vitae

Haomin Chen is completing his Ph.D. degree in Computer Science at the Johns Hopkins University, advised by Assistant Professor Mathias Unberath and Mandell Bellmore Professor Gregory D. Hager. He has been interested in computer vision and machine learning, especially dedicated to interpretable machine learning in medical image analysis during his Ph.D. research. Before that, he obtained a Master of Arts and Science degree in Statistics from Columbia University in 2017, and a Bachelor of Science degree in Physics from Fudan University in 2016. He was introduced to computer vision and machine learning when he was a second-year master's student and has been gradually fascinated by the magical research world since then. As a Ph.D. student, he was devoted to the human-centered design of interpretable machine learning for medical image analysis, founded by the Emerson Collective Cancer Research Fund. He interned in Meta, NVIDIA, PAII, and PingAn Technology, where he was very fortunate to work with so many amazing and professional people from the industry.