# COMPUTATIONAL PATHOLOGY FROM MICROSCOPE TO SOCIAL MEDIA

A Dissertation
Presented to the Faculty of the Weill Cornell Graduate School
of Medical Sciences
in Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy

by
Andrew J. Schaumberg
May 2020

© 2020 Andrew J. Schaumberg

# COMPUTATIONAL PATHOLOGY FROM MICROSCOPE TO SOCIAL MEDIA

Andrew J. Schaumberg, Ph.D.
Cornell University 2020

Pathologists must rapidly interpret microscopic images of a tissue specimen to provide a diagnosis. To diagnose, a pathologist may additionally consider or interpret patient history, radiology, specialized histological stains, genetic sequencing, etc. Integrating into an accurate predictive model each patient's diverse information is one of many challenges in computational pathology. This thesis' scope is computational pathology methods that seek to empower pathologists by automating tedious work, providing new capabilities, or finding similar patient cases. First, we video-recorded pathologists diagnosing at a microscope, and found deep learning could accurately predict where pathologist gaze would dwell. Such areas may be priority regions of interest for diagnosis. Second, we applied deep learning to prostate cancer whole slide images to predict if a cancerous tumor has a SPOP gene mutation. Such methods may expand the capabilities of hospitals where genomic sequencing is unavailable. Third, we integrated histological imaging and clinical information into a multimodal method to find similar patient cases across social media and notify the pathologists having these cases. Such a method brings to low-resource hospitals the subspecialty expertise of pathologists worldwide. Taken together, we conclude it is practical for such computational methods to empower pathologists on an international scale. In doing so, we were the first to predict a somatic mutation from histology images of prostate cancer, we uncovered the utility of pathology data on social media, and we devised novel methods to interpret deep learning.

# BIOGRAPHICAL SKETCH

While an undergraduate at the University of Wisconsin-Platteville [UWP], Andrew J. Schaumberg performed a 7-month work-study as a Co-op Controls Engineer at LasX Industries Inc and a 3-month research fellowship at the National Institute of Standards and Technology [NIST]. Terrie Wheeler mentored Andrew at NIST, with a project to study how the Millennial generation uses information and the NIST library for research. Because this study's survey suggested entering the same search query into many literature databases was problematic for Millennials, Andrew implemented a full-text federated search engine over open-access databases with Apache Lucene, OAI-PMH, and Ruby on Rails.

Andrew graduated from UWP with honors in 2007, with a major in Software Engineering and a minor in Mathematics. After working in software engineering and high performance computing at Cray Inc, National Jewish Health, and the US Department of Energy's Joint Genome Institute, Andrew returned to academia. He pursued a PhD in the Tri-Institutional training program in Computational Biology and Medicine, across Cornell University, Weill Cornell, and Memorial Sloan Kettering. His thesis mentor was Thomas Fuchs, and F31 grant sponsor was Mark Rubin. Andrew reunited with Terrie, who now directed Weill Cornell's library. At this time, Andrew implemented a multimodal text and image search tool over pathology cases on social media and PubMed.

Terrie transformed Weill's library to focus on study space and tools, e.g. 3D printing. Andrew learned to use the library's 3D printer, then applied these skills during the COVID-19 pandemic. At the time of this writing, he prints and freely distributes face masks to people in New York City [NYC], face shields to frontline healthcare workers in NYC, and telepathology equipment to pathologists internationally. See `http://pathobotology.org` for more information.

*In loving memory of Nestor John Grabowski, and to all patients.*

# ACKNOWLEDGEMENTS

I am grateful to many, without whom this thesis would not have been possible.

First, I would like to thank my thesis mentor Prof Thomas Fuchs for introducing me to the field of computational pathology and the diverse challenges of the field. For mentoring me closely at first, but encouraging independence by having me convince myself. For connecting me with institutional pathologists early on, and clearing the way for me to pursue social media projects, which connected me to pathologists worldwide. For teaching me deep learning techniques, while recommending simpler baseline methods for comparison. For imparting to me that statistics and quantitative analyses are not enough, that findings must “make sense”.

I would like to thank my F31 grant sponsor Prof Mark Rubin for inspiring my interest in histological phenotypes of SPOP mutation in prostate cancer. For reviewing numerous slide images with me. For years of meetings, teaching me more about cancer metabolism, signal transduction, and therapeutics. For pathology-related thesis feedback. I thank Rubin lab, Xiaoai Chen, and Elizabeth Manning for extraordinary help with the grant.

I am grateful to Prof David Christini for important guidance and support. Both before I had chosen a thesis laboratory, and as my PhD progressed, I benefited from his advice whenever academic challenges arose. I am also grateful for the feedback from my committee, Prof Christina Leslie, Prof Michael Berger, and Dr Joseph Sirintrapun. I thank Prof Leslie for thesis feedback. I thank Dr Sirintrapun for taking additional time to teach me pathology at the microscope, for being there for my mentees, and for answering multifarious questions during the social media project’s early stages.

I thank Prof Mariam Aly for her supervision of the social media project. For

patiently taking months to convince me to be on social media. For her human subjects research expertise, and ensuring the Institutional Review Board was in the loop before we downloaded any data. For teaching me how to communicate better on social media and with my mentees. For years of detailed feedback on my analyses, writing, and reviews. For vigorously supporting our efforts to help others affected by the pandemic, and in general making me a better person.

I am grateful to numerous pathologists worldwide, who selflessly help others through social media, including me. I thank Dr Laura Pastrían for her leadership of our project, for helping to keep everyone in contact, and for being so approachable when I was new. Dr Mario Prieto Pozuelo for fluorescence *in situ* hybridization guidance, and for being first to share cases with our project. Dr Aurélien Morini for detailed case reviews, highlighting disease state differences among images in a single social media post. Dr Bobbi Pritt for parasitology guidance. For weeks-long formative discussion among these pathologists – and Drs Nusrat Zahra, Duygu Betel Sener, Stephen Yip, Bin Xu, Srinivas Rao Annavarapu, and Kathia Rosado-Orozco – which defined our notion of "disease state" of patient cases posted to social media. Dr Jerad Gardner for later providing corroborating literature of this notion. Dr Sanjay Mukhopadhyay for a phone call that early on convinced me of the clinical importance of the "benign vs malignant" question, and for helpful feedback thereafter. Dr Olaleke Folaranmi for the importance of white balance of photomicrographs. Dr Celina Stayerman for demonstrating extraordinary attention to detail in histopathology cases, and motivating future projects. I thank all pathologist collaborators for sharing your ideas and cases. This thesis is largely built on your goodwill.

I thank my family, for patiently supporting me for decades, throughout my remote work.

# TABLE OF CONTENTS

Biographical Sketch . . . . . . . . . . . . . . . . . . . . . . . . . . . iii
Acknowledgements . . . . . . . . . . . . . . . . . . . . . . . . . . . v
List of Tables . . . . . . . . . . . . . . . . . . . . . . . . . . . . . x

**1 Introduction 1**
1.1 Overview and motivation . . . . . . . . . . . . . . . . . . . . 1
1.2 Research development . . . . . . . . . . . . . . . . . . . . . 2
1.2.1 Microscopy slide region of interest prediction . . . . . . . 3
1.2.2 Mutation prediction from cancer histopathology . . . . . . 4
1.2.3 Connecting pathologists through patient case search . . . 5
1.3 Prevalent themes . . . . . . . . . . . . . . . . . . . . . . . . 8
1.3.1 Obtaining labeled data of sufficient quantity and diversity 8
1.3.2 Interaction, iteration, and multimodality matter . . . . . . 12

**2 Microscopy slide region of interest prediction from video recordings 17**
2.1 Scientific background . . . . . . . . . . . . . . . . . . . . . . 17
2.2 Materials and methods . . . . . . . . . . . . . . . . . . . . . 19
2.2.1 Pathologists . . . . . . . . . . . . . . . . . . . . . . . . . 19
2.2.2 Patient slides . . . . . . . . . . . . . . . . . . . . . . . . 19
2.2.3 Scan preprocessing . . . . . . . . . . . . . . . . . . . . . 20
2.2.4 Video acquisition . . . . . . . . . . . . . . . . . . . . . . 21
2.2.5 Camera choice . . . . . . . . . . . . . . . . . . . . . . . 21
2.2.6 Video preprocessing and registration . . . . . . . . . . . 23
2.2.7 Deep learning . . . . . . . . . . . . . . . . . . . . . . . . 27
2.2.8 Slide magnifications and spatial partitions . . . . . . . . 28
2.2.9 Viewing time threshold for positive/negative ground truth 28
2.2.10 Data augmentation . . . . . . . . . . . . . . . . . . . . . 30
2.2.11 Cross-validation and testing . . . . . . . . . . . . . . . . 31
2.3 Results . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 33
2.4 Conclusion . . . . . . . . . . . . . . . . . . . . . . . . . . . . 37

**3 SPOP mutation prediction from prostate cancer histopathology 39**
3.1 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . 39
3.2 Results . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 44
3.2.1 Molecular information as labels of pathology images opens a new field of molecular pathology . . . . . . . . . . 44
3.2.2 SPOP mutation state prediction is learnable from a small set of whole slides stained with hematoxylin and eosin . . 46
3.2.3 SPOP mutation state prediction is accurate and the prediction uncertainty is bounded within a confidence interval 47
3.2.4 SPOP mutation state prediction is automated and does not rely on human interpretation . . . . . . . . . . . . . . . 47

3.2.5 SPOP mutation state prediction finds patients that appear to share the same genetic driver of disease . . . . . . . . . 48
3.2.6 Molecular pathology, such as characterizing histology in terms of SPOP mutation state, leads directly to precision medicine . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 49
3.2.7 SPOP mutation state prediction provides information regarding other molecular states . . . . . . . . . . . . . . . . . 49
3.3 Discussion . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 50
3.4 Materials and methods . . . . . . . . . . . . . . . . . . . . . . 52
3.4.1 Ensemble aggregation . . . . . . . . . . . . . . . . . . . . 52
3.4.2 Independent evaluation . . . . . . . . . . . . . . . . . . . 53
3.4.3 CBIR evaluation . . . . . . . . . . . . . . . . . . . . . . . 54
S3.5 Supplementary materials and methods . . . . . . . . . . . . . . 55
S3.5.1 Digital slide acquisition and quality control . . . . . . . . . 55
S3.5.2 Region of interest identification . . . . . . . . . . . . . . . 56
S3.5.3 Data augmentation . . . . . . . . . . . . . . . . . . . . . . 56
S3.5.4 Deep learning . . . . . . . . . . . . . . . . . . . . . . . . . 57
S3.5.5 Cross-validation . . . . . . . . . . . . . . . . . . . . . . . 57
S3.5.6 TCGA dataset learning . . . . . . . . . . . . . . . . . . . . 61
S3.5.7 MSK-IMPACT dataset testing . . . . . . . . . . . . . . . . 62
S3.5.8 Implementation details . . . . . . . . . . . . . . . . . . . . 65

**4 Connecting pathologists through patient case search on social media and PubMed 73**
4.1 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . . 74
4.2 Materials and methods . . . . . . . . . . . . . . . . . . . . . . 78
4.2.1 Social media data . . . . . . . . . . . . . . . . . . . . . . 78
4.2.2 PubMed data . . . . . . . . . . . . . . . . . . . . . . . . . 80
4.2.3 Image processing . . . . . . . . . . . . . . . . . . . . . . . 80
4.2.4 Text processing . . . . . . . . . . . . . . . . . . . . . . . . 81
4.2.5 Random Forest classifier . . . . . . . . . . . . . . . . . . . 81
4.2.6 Customized hybrid deep-learning-random-forest model and clinical covariates . . . . . . . . . . . . . . . . . . . . . 82
4.2.7 Disease state classifier repurposed for similarity-based search . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 84
4.2.8 Three levels of sanity checking for search . . . . . . . . . . 85
4.2.9 Five levels of method interpretability . . . . . . . . . . . . 86
4.2.10 Experimental design and evaluation . . . . . . . . . . . . . 91
4.2.11 Social media bot for public prospective testing . . . . . . . 93
4.2.12 Computational hardware . . . . . . . . . . . . . . . . . . . 93
4.3 Results . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 94
4.3.1 Identifying and filtering for H&E images . . . . . . . . . . 94
4.3.2 Distinguishing common stain types . . . . . . . . . . . . . 95
4.3.3 Distinguishing ten histopathology tissue types . . . . . . . 96

4.3.4 Deep learning predicts disease state across many tissue types . . . 98
4.3.5 Texture and tissue are important clinico-visual features of disease . . . 98
4.3.6 Disease state search, first pan-tissue pan-disease method . 100
4.4 Discussion . . . 102
4.4.1 Summary . . . 102
4.4.2 Comparison with prior studies . . . 104
4.4.3 Caveats and future directions . . . 105
S4.5 Supporting information . . . 109
S4.5.1 Image data overview . . . 109
S4.5.2 Supplementary Image processing . . . 113
S4.5.3 Data diversity discussion . . . 118
S4.5.4 Procedure overview . . . 119
S4.5.5 Text data overview . . . 124
S4.5.6 Supplementary Text processing . . . 126
S4.5.7 Nontumoral, Low grade, and Malignant task details . . . 128
S4.5.8 Image features for machine learning . . . 138
S4.5.9 Machine learning sanity checking for search . . . 141
S4.5.10 Machine learning interpretability for search . . . 143
S4.5.11 Machine learning methods discussion . . . 151
S4.5.12 Disease state search distance function and performance . . 158
S4.5.13 Supplementary Disease state search results . . . 161
S4.5.14 Supplementary Experimental design and evaluation . . . 171
S4.5.15 Supplementary Computational hardware and software . . 171
S4.5.16 Supplementary Comparison with prior studies . . . 172
S4.5.17 Supplementary Caveats . . . 176
S4.5.18 Supplementary Future directions . . . 179

**5 Conclusion** **184**

# LIST OF TABLES

2.1 Accuracies of ten trials of three-fold cross-validation in bladder. . 34
2.2 Accuracies of ten trials of three-fold cross-validation in prostate. 35

S3.1 TCGA test set images and accuracies . . . . . . . . . . . . . . . . . 67
S3.2 TCGA test set images and accuracies, continued . . . . . . . . . . 68

S4.1 Case similarity search performance in detail. . . . . . . . . . . . . 159
S4.2 Case similarity search performance per-pathologist. . . . . . . . . 161

# LIST OF ALGORITHMS

2.1 Automated registration procedure to find the least off-center whole slide image patch from a camera video frame. . . . . . . . . 24

3.1 Dominant tumor cell counting as slide preprocessing . . . . . . . . 58
3.2 Metaensemble training. . . . . . . . . . . . . . . . . . . . . . . . . . . . 59
3.3 Metaensemble SPOP mutation prediction. . . . . . . . . . . . . . . . . 60

## LIST OF FIGURES

1.1 Random Forest classification and proximity schematics. . . . . . 6
1.2 Random Forest interpretability of deep features. . . . . . . . . . . 7

2.1 Proposed microscope-based saliency predictor pipeline workflow. 18
2.2 Slides and tissues. . . . . . . . . . . . . . . . . . . . . . . . . . 20
2.3 Optical flow, showing pixel movement grid. . . . . . . . . . . . 20
2.4 Image registration. . . . . . . . . . . . . . . . . . . . . . . . . . 23
2.5 Lens change detection. . . . . . . . . . . . . . . . . . . . . . . . 25
2.6 CaffeNet architecture. . . . . . . . . . . . . . . . . . . . . . . . 27
2.7 Pathologist viewing times in seconds at the microscope . . . . . . 28
2.8 Scaled image patches of left and right sides of bladder patient slide. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 29
2.9 Ten three-fold cross-validation trials for bladder and prostate. . . 30
2.10 Interpatient Area Under the Receiver Operating Characteristic for bladder and prostate . . . . . . . . . . . . . . . . . . . . . . 33

3.1 TCGA and MSK-IMPACT slides for SPOP study. . . . . . . . . . 40
3.2 SPOP mutations in the Interated Genomics Viewer . . . . . . . . 41
3.3 SPOP mutation prediction pipeline . . . . . . . . . . . . . . . . 42
3.4 Deep learning architecture and activation heatmaps . . . . . . . . 43
3.5 SPOP prediction performance. . . . . . . . . . . . . . . . . . . . 45
3.6 CBIR performance . . . . . . . . . . . . . . . . . . . . . . . . . 48
S3.7 SPOP teritary structure with mutated residues. . . . . . . . . . . 65
S3.8 TCGA classifier and ensemble mutation prediction scores. . . . . 66
S3.9 Ensemble performance . . . . . . . . . . . . . . . . . . . . . . . 69
S3.10 Tuned ensemble performance . . . . . . . . . . . . . . . . . . . 70

4.1 Graphical summary. . . . . . . . . . . . . . . . . . . . . . . . . 76
4.2 Technique, tissue, and disease diversity. . . . . . . . . . . . . . . 79
4.3 Deep learning methods summary. . . . . . . . . . . . . . . . . . 82
4.4 Random Forest feature importance for prioritizing deep features, when non-deep, deep, and clinical features are used together for learning. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 87
4.5 Interpretable spatial distribution of deep learning predictions and features. . . . . . . . . . . . . . . . . . . . . . . . . . . . . 89
4.6 Disease state clusters based on hand-engineered, natural-image-trained deep features, or histopathology-trained deep features. . 92
4.7 H&E performance. . . . . . . . . . . . . . . . . . . . . . . . . . 94
4.8 10-tissue type and 3-disease state prediction performance and counts. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 96
4.9 Disease state prediction performance for machine learning methods. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 97
4.10 Case similarity search performance. . . . . . . . . . . . . . . . . 99

S4.11 Artifact and foreign body images. . . . . . . . . . . . . . . . . . . . 110
S4.12 Parasitology images. . . . . . . . . . . . . . . . . . . . . . . . . . 110
S4.13 Gross images. . . . . . . . . . . . . . . . . . . . . . . . . . . . . 111
S4.14 Image acceptability criteria. . . . . . . . . . . . . . . . . . . . . . 111
S4.15 H&E images. . . . . . . . . . . . . . . . . . . . . . . . . . . . . 112
S4.16 Interactive Pathology Annotator tool and social media dialogue. 121
S4.17 Interactive Pathology Annotator tool and tissue type hashtags. . 122
S4.18 3-disease text processing algorithm flowchart. . . . . . . . . . . 125
S4.19 Machine learning features. . . . . . . . . . . . . . . . . . . . . . 139
S4.20 Random Forest feature importance for prioritizing only-natural-image-trained deep features, when non-deep, ImageNet$_{2048}$, and clinical features are used together for learning. . . . . . . . . . . . 145
S4.21 Interpretable spatial distribution of deep learning predictions and features of multiple cases. . . . . . . . . . . . . . . . . . . . . . 149
S4.22 Disease state clusters after dimensionality reduction. . . . . . . . 150
S4.23 Deep learning performance curves. . . . . . . . . . . . . . . . . . 152
S4.24 Deep features. . . . . . . . . . . . . . . . . . . . . . . . . . . . . 153
S4.25 Sigmoid vs centered soft clipping performance. . . . . . . . . . . 154

CHAPTER 1

# INTRODUCTION

The author makes this chapter available under the terms of a Creative Commons Attribution-NonCommercial-NoDerivatives [CC-BY-NC-ND] 4.0 International license, at `https://creativecommons.org/licenses/by-nc-nd/4.0/`.

## 1.1 Overview and motivation

Computational pathology demands a tremendous amount of labor from many. After undergoing decades of training, a pathologist's clinical service obligations involve diagnosing potentially hundreds of patient cases daily and entering this into medical records. Additionally, laboratory staff meticulously follow protocol to prepare glass microscopy slides for pathologists, sequence genomes, and perform other tests as needed. Technicians digitally scan glass microscopy slides whole, archive glass slides, store slide images, and maintain supercomputers that run artificial-intelligence-driven computational models to process these images. Engineers continuously improve all the instruments used, ranging from the clinical laboratory to the datacenter. Informaticists conceive new computational models to gain insight into pathological processes, often using data from medical records, genomic sequencing, and whole slide images. All of this is done in the interests of the patient, who typically must take the time to come to the hospital and submit a tissue sample, in coordination with administrative staff, primary care providers, and other specialists.

It would be desirable for all this work to make the world a better place, but where might one read a plan for how to make the world a better place? We observe the United Nations' Sustainable Development Goals outline a 17-point plan to be a "blueprint to achieve a better and more sustainable future for all". Sustainable Development Goal 3: Good Health and Well-Being suggests "ensuring healthy lives and promoting the well-being at all ages is essential" and "increased access to physicians" is important[1]. To this end, it is recommended to "[s]ubstantially increase [...] the recruitment, development, training, and retention of the health workforce in developing countries" to achieve "access to quality essential health care services" for all[2]. As a way to increase access to physicians and quality health care services in general, we focus on connecting pathologists worldwide through computational pathology tools. To us, a world of connected pathologists means it is easier for pathologists to immediately obtain second opinions from other pathologists anywhere in the world. This connectivity would broaden consensus on the next steps of any patient's care, potentially confronting even the time pressures of diagnosis during surgery.

## 1.2 Research development

**Timeline and summary** Research began with a pilot study of pathologists making diagnoses at a microscope, which is detailed in chapter 2. Some chapter 2 work was repurposed to identify putative regions of interest for predicting a gene mutation in prostate cancer, which is detailed in chapter 3. This effort continues at the time of this writing. To make a more robust search/CBIR tool than was possible in chapter 3, this thesis concludes with a study of pathologists worldwide, to train an artificial intelligence [AI] that can find similar patient cases on social media (see chapter 4). A recurring challenge in these computational pathology projects is a "needle in a haystack problem". Namely, given a whole microscopy slide, it is a challenging for an AI to find a few important regions of interest to solve a problem, whether the problem is to predict the patient's disease or to predict a malignant tumor's underlying driver mutation.

### 1.2.1 Microscopy slide region of interest prediction

**Challenging for AI to identify regions of interest in a whole slide image** Computational pathology tasks often have a fundamental "needle in a haystack problem", where a patient's glass microscopy slide (i.e. the "haystack") may be summarized by a few select regions of interest (i.e. the "needle"), but the locations of these regions are not explicitly written down for an AI to know from the beginning. Moreover, a pathologist may examine multiple slides of a patient and share only one region of interest with pathologist colleagues on social media to ask for second opinions[3]. How might an AI learn to identify where regions of interest are in any slide?

**Video-recording is one of many possible ways to address challenge** Indeed, our second Strategic Aim sought to cross-reference a whole slide image with a video recording of a pathologist making a diagnosis at a microscope. If a pathologist spent more than 0.1 seconds looking at a region, then that region was deemed of interest[4][1]. In a pilot study of two pathologists for prostate and bladder cancers, we trained a Convolutional Neural Network [CNN] to accurately predict if a region is of interest to a pathologist making a diagnosis at a microscope[4]. By separating the "needle" of the pathology from the rest of the "haystack" of the slide, downstream computational methods may be better able to (i) identify disease in slides, (ii) draw pathologist attention to regions of interest to save the pathologist time, and (iii) efficiently search through only regions of interest in large digitized whole slide libraries.

---

[1] The assumption that pathologists will look longer at regions of interest is not without caveats. One may suggest instead pathologist need only a brief time to immediately recognize diagnostically important regions, while longer looks may indicate a pathologist's curiosity. We leave to future work comparisons of pathologist view timings and a pathologist's explicit labels of diagnostic importance.

### 1.2.2 Mutation prediction from cancer histopathology

**Predicting SPOP mutation from whole slide images without pathologist intervention** The presence or absence of "driver mutations" in a patient's cancer is an alternative way to label microscopy images for an AI to learn, and this does not require a pathologist's diagnosis or video-recorded behavior. We studied the SPOP gene in prostate cancer, which is mutated in ~10% of cases[5]. Our method identified regions of interest on the basis of atypical nuclei density. For these regions we trained ensembles of CNNs to (i) predict if the SPOP gene was mutated in this patient's cancer tumor and (ii) estimate the uncertainty in this prediction, e.g. our method may estimate that a particular patient has a 33-56% chance of having an in-tumor SPOP mutation.

**Challenges in this genomic approach** However, predicting underlying mutations that drive malignant growth in a tumor is difficult because some mutations may not be histologically evident, some mutations may not be histologically distinguishable from other mutations, and some mutations may be so uncommon that machine learning is not feasible for such few patient cases. Moreover, between institutions the clinical protocols may differ in ways that fundamentally change slide appearance, e.g. microscopy slides are typically prepared as frozen sections in the TCGA patient cohort while microscopy slides are prepared as formalin-fixed paraffin-embedded [FFPE] sections in the Memorial Sloan Kettering Cancer Center patient cohort. Freezing damages and tears tissue, but this does not occur with FFPE sections. It follows that an AI trained on FFPE section slides may become inaccurate when the AI is tested on frozen section slides[5].

**Broader impacts of this work** Nonetheless, our work is widely cited by pathology-focused[6,7,8,9,10] and computationally-focused[11,12,13,14,15] literature in the field. We believe our ongoing work is the first to predict a somatic mutation in prostate cancer from H&E-stained whole slide images alone.

### 1.2.3 Connecting pathologists through patient case search

**If a patient case is challenging, find similar cases to cross-reference** In light of this priority to connect pathologists worldwide, it follows that the the grant that largely funded this research would involve *search*, which is formally referred to as “content-based image retrieval” [CBIR][3]. Our originally-proposed search method sought to find similar patient cases in The Cancer Genome Atlas [TCGA] patient cohort, given photomicrographs from a textbook or microscope. Our rationale is a pathologist considering a challenging patient case will be interested in similar patient cases, and automated *search* may find similar cases.

**Search connects pathologists who shared similar cases on social media** We later realized photomicrographs of de-identified patient cases are frequently shared by pathologists through social media and the published literature, so

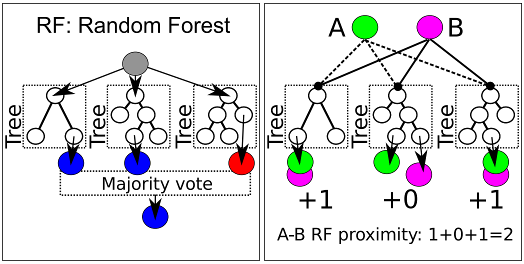


Figure 1.1: **Random Forest classification and proximity schematics.** A Random Forest [RF] consists of many Random Trees, e.g. three trees, but typically 1,000 trees are used. *At left* a RF classifies a datum (*gray circle*) according to a majority vote of the trees, e.g. two votes for the "blue class" and one vote for the "red class" means the datum is classified as "blue". Broadly, for our purposes, each datum is an image, and the Random Forest must classify the image's depicted disease state (i.e. nontumor, low grade, or malignant)[3]. *At right* an RF is reused for *search*, were two data are considered similar if they fall into the same leaf of a tree. For example, if images A and B are in the same leaf for two of the three trees in the RF, then the A-and-B RF similarity/proximity is 2.

we implemented search tools specifically for these media[16,3]. Powerfully, the connectivity of social media allows a pathologist to not only find similar cases through *search*, but also *confer* with pathologists who had similar cases. Our social-media-focused approach to has been well-received in both the search-focused[17] and pathology-focused[18,19,20] literature as our methods developed over time. Social media has offered a tractable way for us to leverage machine learning to serve, connect, and study pathologists worldwide. Ultimately, we implemented an artificial-intelligence-driven social media bot, to both find pathologists having similar cases and connect these pathologists in a conversation to discuss the next steps in a patient's care[3].

**Method interpretability suggests which features are important** Our social media bot's underlying artificial intelligence [AI] is a machine learning classifier trained to predict disease state, but this classifier is reused for search (Fig 1.1). Additionally, this classifier leverages hand-engineered features and shallow

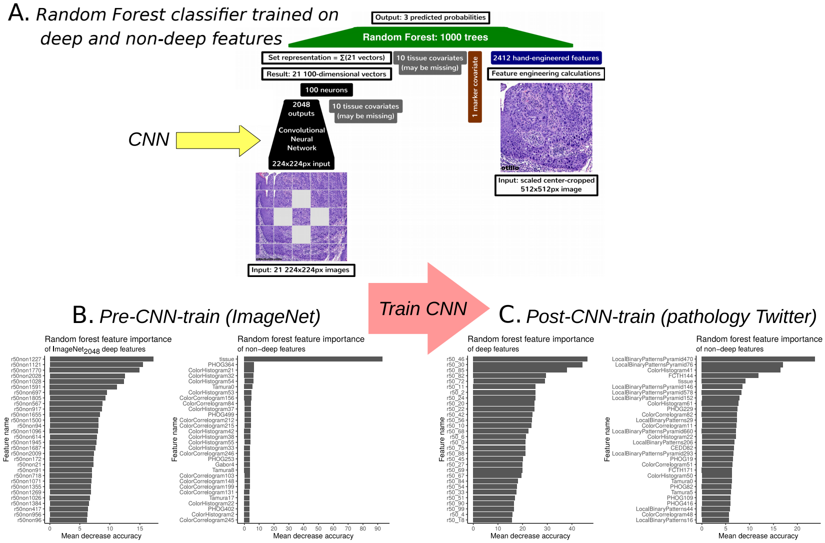


Figure 1.2: **Random Forest interpretability of deep features.** **(A)** We combine deep learning and shallow learning techniques to accurately classify disease state for patient cases posted to social media[3] (c.f. Fig 4.3). Both before **(B)** and after **(C)** training on patient case data (c.f. Figs 4.4 and S4.20, respectively), we measure Random Forest feature importance to understand what concepts the deep features from a convolutional neural network [CNN] have learned to represent. From the before-and-after changes in Random Forest feature importance one may infer what deep CNN features learned to represent from the patient case data.

learning (i.e. a Random Forest) to interpret deep learning (i.e. a Convolutional Neural Network [CNN]) features (Fig 1.2). In this way, we find tissue type (e.g. breast, dermatological, gastrointestinal, etc), color, and edge features are *less* important after CNN training. However, texture features are *more* important after CNN training. Therefore, as further explained in our recent work[3], we reason that the CNN learns histopathology-relevant edge, color, and tissue features from patient cases on social media (which reduces the importance of e.g. hand-engineered edge and color features after learning), but the deep neural network may forget histopathology-relevant texture features during learning (which increases the importance of hand-engineered texture features after learning). It "makes sense" that the CNN learns tissue type is important, e.g. infectious disease is more common in gastrointestinal tissue than breast tissue. When the concepts learned make sense broadly, one becomes confident that the AI has learned to perform a task in ways that make sense too.

## 1.3 Prevalent themes

### 1.3.1 Obtaining labeled data of sufficient quantity and diversity

For all strategic aims, we found it is challenging to obtain labeled images in sufficient quantity and diversity for accurate machine learning in computational pathology.

#### Obtaining labeled images

**Where a disease is in a slide is often not written down** Obtaining labeled images is a challenge because pathologists are often required to write down what disease is shown in a small image focused on that disease morphology. Moreover, because whole slide images are so large, the "needle in a haystack problem" arises if all that is written down is a patient's disease, without any localization of the disease in the slide. The needle in a haystack problem remains for bulk genomic sequencing, e.g. chapter 3, where a machine measures the presence or absence of a genetic mutation in a tumor, but we do not know exactly where in a slide the morphological phenotype for this mutation is evident[5].

**Crowd-source among pathologists globally to get more specifics written down** We found the most success from crowd-sourcing image labeling work across pathologists and other collaborators. For chapter 4, social media gave our project worldwide reach to collaborate with pathologists, and mentoring opportunities afforded us the help of talented High School students to curate data[3]. For chapter 2, video-recording pathologists did not disrupt the pathologist's diagnostic workflow, and we did obtain localized timing measurements for all regions throughout a whole slide image[4]. However for chapter 3, skipping a pathologist entirely by relying on genomic sequencing to label the images left us with both the aforementioned "needle in a haystack problem" and an uncertain amount of error from genomic sequencing and tumor purity[5].

**Obtaining a sufficient quantity of images**

**Patient case data in abundance on social media but even more in published case reports** For machine learning generally, more data are always better. This was especially challenging for chapter 4, because any one pathologist shares on social media several to hundreds of labeled photomicrographs, but we believed our machine learning method needed thousands to robustly predict disease state[3]. Moreover, a rare disease may only be seen by a pathologist once in their career, so we believed we needed over 100,000 photomicrographs to have any realistic chance of our search tool (i.e. pathobot) finding rare but similar patient cases. We found that making our search tool freely available and easy to use convinced many pathologists worldwide to share data with us, join our project, and become co-authors of our study. Additionally, public data from manuscripts in PubMed Central were essential to obtain over 100,000 photomicrographs, making our search tool more comprehensive. It was also challenging to filter over 1,000,000 manuscripts in PubMed Central to find over 100,0000 hemotoxylin and eosin [H&E] photomicrographs. However, our machine learning classifier trained to recognize H&E photomicrographs on social media demonstrated accurate performance and was suitable for this task.

**Video recordings provide numerous data across few patient cases** We benefited from having comprehensive pathologist timing information throughout each whole slide studied for chapter 2[4]. Deep learning could effectively be applied to these data for accurate predictions. However, the experimental approach to data collection could not be easily scaled up to many slides or multiple institutions, because a researcher had to physically meet with a pathologist to obtain video recordings. We were not aware of public data that could have further supported the deep learning task in chapter 2.

**Public and private data complementary for genomic studies** In chapter 3 we leveraged public data from TCGA, as well as private data from Memorial Sloan Kettering Cancer Center[5]. Despite the large size of these two patient cohorts and that SPOP is one of the most frequently mutated genes in prostate cancer, we still needed to request additional scans to digitize glass microscopy slides for patient cases which had our mutation of interest (i.e. SPOP). SPOP mutation occurs in ~10% of prostate cancer patient cases, so for 1000 patients only 100 typically have SPOP mutation. There are logistical challenges to involve additional private cohorts, and we are not aware of other public cohorts.

### Obtaining a sufficient diversity of images

**Patient case data from social media highly diverse** For chapter 4, the pathology images from social media are almost too diverse[3]. Pathologists may take photomicrographs with a microscope-mounted camera or smartphone. Laboratory and staining protocols differ between institutions and countries, e.g. saffron may be used in H&E sections in France[3]. Pathologists focus on different organ systems, each with their own distribution of diseases. For instance, infectious disease is more common in gastrointestinal and pulmonary tissues than neurological or breast tissues. Pathologists have different microscopes, and may illuminate their samples differently. In light of such diversity, some image normalization may be helpful, e.g. white balance correction. Importantly, such diversity among the numerous pathologists participating in a study allows computational methods to be rigorously tested through a "leave one pathologist out" cross-validation approach. Through "leave one pathologist out" testing, an AI may be trained on highly diverse data from all pathologists except one. AI performance is then tested on this one held-out pathologist's data. This procedure is repeated, with each pathologist held-out in turn for testing, and the test performance results are averaged to estimate the AI's performance in general for any pathologist.

**Video recordings across few patient cases not diverse** For chapter 2, the DeepScope data are likely not diverse enough[4]. These data came from a single institution, so it is not possible to estimate how the method performs in general across institutions. However, it was possible to estimate performance of an AI trained on one tissue (e.g. bladder cancer) tested on a different tissue (e.g. prostate cancer).

**Histopathology-genomic paired data from different patient cohorts may qualitatively differ** For chapter 3, the SPOP data are diverse in a difficult-to-compare manner[5]. TCGA data are frozen sections. Memorial Sloan Kettering Cancer Center data are FFPE sections. Therefore, when measuring performance when training on one cohort and testing on the other cohort, it may not be clear if performance changes are due to differences in sectioning protocol. Including more cohorts for a ”leave one cohort out” approach was not logistically possible.

## 1.3.2 Interaction, iteration, and multimodality matter

### Interaction

**Interaction facilitates crowd-sourcing** Computationally, we found the most success by making tools pathologists can readily use anywhere worldwide for free[3]. Smartphones and social media are ubiquitous. The vast majority of the time, a pathologist would use our tools by mentioning “@pathobot” in a social media post, which is 10 of the maximum 280 characters in a Twitter post. This accessible user interface allowed our tools to be seamlessly embedded in the conversations of pathologists worldwide, which facilitated further recruitment of pathologists to our chapter 4 project.

**Interaction facilitates consensus** In pathology, interaction with colleagues is an important way to establish consensus. For instance, some of our pathologist collaborators reported working alone. Social media offers a way to compare one's own opinion against the opinions of other pathologists. Further, interaction allows a pathologist to use our search tools numerous times for the same patient case, including different photomicrographs, additional diagnostic terms, specific morphological descriptions, or special commands each time until the desired results are obtained.

### Iteration

**Iterate for improved performance and expanded case database** Each time a pathologist was recruited through pathobot, we curated and integrated their data into our patient case database[3]. This occurred frequently enough that we developed the Integrated Pathology Annotator [IPA] tool to help pathologists and other researchers to review the correctness of data annotations. IPA also automatically regenerates all database files for machine learning. From these updated data files an updated machine learning classifier may be trained. We believe iteratively improved classifier performance and an expanded case database served to further recruit additional pathologists, due in part to social network effects. Anecdotally, we found that some collaborating pathologists observed our rapid incremental recruitment and would recommend specific colleagues to collaborate in our study.

**Iterative consensus and search** In pathology, iteration may be an important part of establishing a diagnosis. Pathologists may use social media to not only ask for opinions to make a diagnosis, but also ask what tests should be ordered next to establish a diagnosis. This incremental winnowing of potential diagnoses from a differential diagnosis may occur though checking patient history, ordering of specific immunohistochemistry stains, etc. Iteration also allows a pathologist to use our search tools repeatedly over time as more information for the case becomes available, or as colleague opinions accumulate for context. In this way, for a specific patient case in question, the search for similar patient cases is crowd-sourced as well, across the aggregating opinions of all colleagues working together to reach consensus for this case.

### Multimodality

To diagnose, a pathologist may consider or interpret many modalities besides H&E, including patient history, radiology, gross sections, specialized histological stains, cytological smears, genomic sequencing, and other laboratory tests.

**Multimodality in ARDS, DAD, and COVID-19** For example, patient history, chest X-rays, computed tomography [CT] scans, and genetic risk factors may be considered for a clinical diagnosis of Acute Respiratory Distress Syndrome [ARDS][21]. Poor survival among patients with ARDS may make ante-mortem biopsies difficult to obtain in time for a pathological diagnosis. The histological hallmark of ARDS is Diffuse Alveolar Damage [DAD], but multiple causes of ARDS and DAD exist[21]. A polymerase chain reaction [PCR] assay on adjacent tissue blocks or nasopharyngeal swabs may elucidate specific causes of ARDS/DAD, such as COVID-19 infection[22]. For COVID-19 patients, H&E sections may indicate whether a patient died *with* COVID-19 or *from* COVID-19[23].

**Multimodality in prostate cancer** A separate example in contrast is prostate cancer, where H&E sections are more frequently available while the patient is still alive. Complementing H&E, genomic sequencing may identify both "indolent" tumors that require no further treatment and "aggressive" tumors driven by specific genetic mutations that may require precision medicine to effectively treat[24]. Though prostate cancer may be diagnosed on the basis of H&E sections, the diagnosis may be further informed by laboratory tests such as serum Prostate Specific Antigen [PSA] levels, or by IHC stains such as p63 or GATA3[24]. Further, clinical/radiological follow-up index of suspicion may be part of patient care follow-up plans[24]. Thus, in these examples, H&E-stained microscopy slides may be interpreted in the context of other modalities to establish diagnosis, guide treatment decisions, and/or determine cause of death.

**Multimodality in our methods** Indeed, a multimodal diagnostic approach is so important that pathologists often share multiple modalities in a single social media post, as reflected in our data[3]. Thus it is essential for computational pathology methods to handle multiple modalities, even if methods such as ours for chapter 4 focus on the two most common modalities: H&E and immunohistochemistry [IHC]. Our methods handle H&E explicitly, by training a machine learning classifier on H&E photomicrographs. IHC is handled as an oversimplified covariate. Additionally, our methods consider tissue type (e.g. breast, dermatological, gastrointestinal, etc), which is clinical information passed on to a pathologist. Further, our hand-engineered and natural-image-trained deep features offer a way to generically compare images in a pathology-agnostic manner. Finally, text-based matching allows patient cases with similar diagnoses (or differential diagnoses) to be matched independently of image data. We intend this multimodal approach to offer the greatest value to pathologists searching for opinions of colleagues having similar patient cases, given as much or as little information for the case as is immediately available.

## CHAPTER 2

## MICROSCOPY SLIDE REGION OF INTEREST PREDICTION FROM VIDEO RECORDINGS

Portions of this chapter first appeared in Schaumberg* et al 2017. Schaumberg et al 2016 was written in collaboration among Andrew J. Schaumberg, Peter J. Schüffler, and Thomas J. Fuchs. S. Joseph Sirintrapun and Hikmat A. Al-Ahmadie reviewed Schaumberg et al 2017. The author makes this chapter available under the terms of a Creative Commons Attribution-NonCommercial-NoDerivatives [CC-BY-NC-ND] 4.0 International license, at `https://creativecommons.org/licenses/by-nc-nd/4.0/`.

### 2.1 Scientific background

Computational pathology[25] relies on training data annotated by human experts on digital images. However, the bulk of a pathologist's daily clinical work remains manual on analog light microscopes. A noninterfering system which translates this abundance of expert knowledge at the microscope into labeled digital image data is desired.

Tracking a pathologist's viewing path along the analyzed tissue slide to detect local image saliency has been previously proposed. These approaches include whole slide images displayed on one or more monitors with an eye-tracker[26], mouse-tracker[27] or viewport-tracker[28,29] – but may suffer confounds including peripheral vision[30], head turning[31], distracting extraneous

---

*Schaumberg A.J., Joseph Sirintrapun S., Al-Ahmadie H.A., Schüffler P.J., Fuchs T.J. (2017) DeepScope: Nonintrusive Whole Slide Saliency Annotation and Prediction from Pathologists at the Microscope. In: Bracciali A., Caravagna G., Gilbert D., Tagliaferri R. (eds) Computational Intelligence Methods for Bioinformatics and Biostatistics. CIBB 2016. Lecture Notes in Computer Science, vol 10477. Springer, Cham

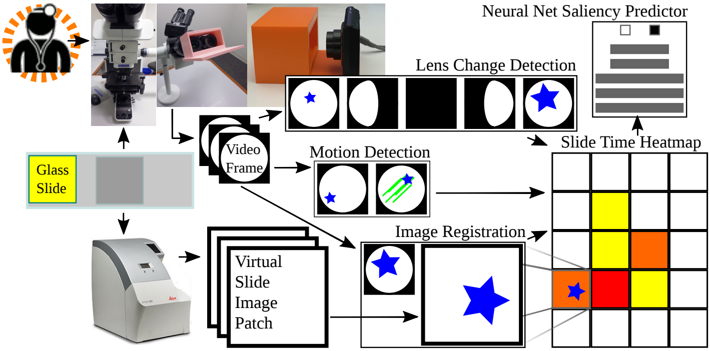

Neural Net Saliency Predictor
Lens Change Detection
Slide Time Heatmap
Video Frame
Motion Detection
Glass Slide
Image Registration
Virtual Slide Image Patch

Figure 2.1: **Proposed microscope-based saliency predictor pipeline workflow.** The pathology session is recorded, the slide is scanned, the video frames are registered to scan patches. Lens change detection guides registration and viewing time is recorded for periods without motion. A convolutional neural net learns to classify patches as salient (long looks) or not.

detail[32], monitor resolution[33], multimonitor curvature[34], and monitor bezel field of view fragmentation[35]. Because computer customizations may potentially effect viewing times, for studies of pathologists recorded at a computer, we suggest noting the sensitivity and choice of pointing device, e.g. trackball, touch pad, touch screen, pointing stick, mouse, and if a scroll wheel or keyboard was available to zoom in or out. Only our approach does not change the pathologist's medical practice from the microscope. The microscope is a class I device appropriate for primary diagnosis according to the United States Food and Drug Administration, while whole slide imaging devices are class III[36].

In light of the confounds of alternatives, its centuries of use in pathology, and its favorable regulatory position for primary diagnosis, we believe the microscope is the gold standard for measuring image region saliency. Indeed, there is prior work annotating regions of interest at the microscope for cytology technicians to automatically position the slide for a pathologist[37].

We therefore propose a new, noninterfering workflow for automated video-based detection of region saliency using pathologist viewing time at the microscope (Fig 2.1). Viewing time is known in the psychology literature to measure attention[38,39], and we define saliency as pathologist attention when making a diagnosis. Using a commodity digital camera, rather than a custom embedded eye-tracking device[40,41], we video record the pathologist's entire field of view at a tandem microscope to obtain slide region viewing times and register these regions to whole slide image scan regions. Second, we train a convolutional neural network [CNN] on these observation times to predict whether or not a whole slide image region is viewed by a pathologist at the microscope for more than 0.1 seconds [s]. As more videos become available, our CNN predicting image saliency may be further trained and improved, through online learning.

## 2.2 Materials and methods

### 2.2.1 Pathologists

Pathologists were assistant attending rank with several years experience each. Trainees have different, less efficient, slide viewing strategies[26,30].

### 2.2.2 Patient slides

Two bladder cancer patients were studied by author SJS. Two prostate cancer patients were studied by author HAA. One slide per patient was used, for four slides total (Fig 2.2).

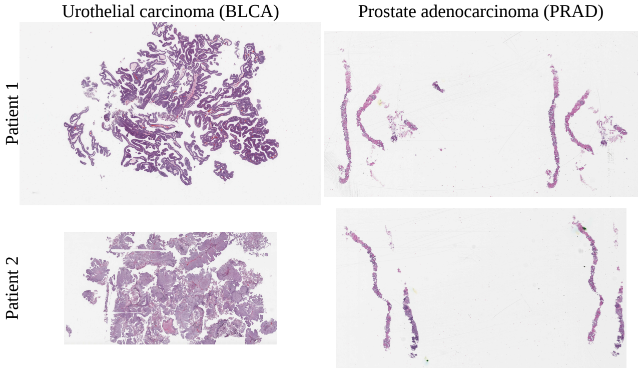


Figure 2.2: **Slides and tissues.** Bladder cancer left, prostate cancer right. Training, validation, testing done on top slides, with additional same-tissue testing on bottom slides. For cross-tissue testing, top slide tested against other top slide. Viewing time heatmap for top left bladder shown in Fig 2.7. Note how the top bladder has more edges than the more solid bottom bladder, while the prostates have similar tissue texture. We believe this impacts interpatient accuracy, shown in Fig 2.9.

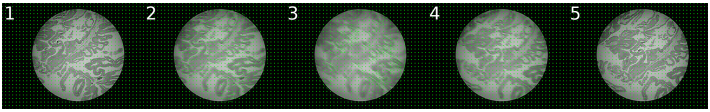


Figure 2.3: **Optical flow, showing pixel movement grid.** The frame has few moving pixels before *(left)* and after *(right)* pathologist moves the slide. A pathologist looks at a slide region for the duration of consecutive stationary frames.

### 2.2.3 Scan preprocessing

Microscope slides, inspected by a pathologist, were scanned at 0.5±0.003 microns per pixel [px], using an Aperio AT2 scanner. The resulting SVS data file consists of multiple levels, where level 0 is not downsampled, level 1 is downsampled by a factor of 4, level 2 by a factor of 16, and level 3 by a factor of 32. From each level, $800 \times 800$ px patches were extracted via the OpenSlide software library[42]. In bladder, adjacent patches in a level overlap at least 50%, to avoid windowing artifacts in registration. In prostate, adjacent patches overlap at least 75%, to best center the pathologist's field of view on the little tissue in a needle biopsy. Patches evenly cover the entire level without gaps. Scans were either taken before a technician applied marker to the slides, to indicate regions of interest to the pathologist, or after markings were scrubbed from the slide. However, these marks were evident in the pathologist videos discussed in the next section.

### 2.2.4 Video acquisition

A Panasonic Lumix DMC-FH10 camera with a 16.1 megapixel charge-coupled device [CCD], capable of 720p motion JPEG video at 30 frames per second fps], was mounted on a second head of an Olympus BX53F multihead teaching microscope to record the pathologist's slide inspection. Microscope objective lens magnifications were 4x, 10x, 20x, 40x, and 100x. Eyepiece lens magnifications was 10x. The pathologist was told to ignore the device and person recording video at the microscope during inspection. The mount (Fig 2.1) for this camera was designed in OpenSCAD and 3D-printed on a MakerBot 2 using polylactic acid [PLA] filament.

### 2.2.5 Camera choice

Many expensive microscope-mounted cameras exist, such as the Lumenera INFINITY-HD and Olympus DP27, which have very good picture quality and frame rate. The Lumenera INFINITY-HD is a CMOS camera, not CCD, so slide movement will skew the image rather than blur it, and we did not want to confound image registration or motion detection with rolling shutter skew. Both cameras trim the field of view to a center-most rectangle for viewing on a computer monitor, which is a loss of information, and we instead assign viewing time to the entire pathologist-viewed 800 × 800 px PNG patch from the SVS file representing the whole slide scan image. Both cameras do not have USB or Ethernet ports carrying a video feed accessible as a webcam, for registration to the whole slide scan. The Olympus DP27 may be accessible as a Windows TWAIN device, but we could not make this work in Linux. Finally, the HDMI port on both carries high-quality but encrypted video information that we cannot record, and we did not wish to buy a Hauppauge HDMI recording device, because we had a cheaper commodity camera on hand already. We also considered automated screenshots of the video feed in Aperio ImageScope as displayed on a computer monitor, but we observed a lower frame rate and detecting lens change is complicated because the entire field of view is not available. Recording low-quality video on a commodity camera to a SecureDigital [SD] memory card is inexpensive, captures the entire field of view, and is generally applicable in any hospital. For this pilot study, we used only one camera for video recording, rather than two different microscope cameras, potentially eliminating a confound for how many pixels are moving during rapid short movements of the slide. For 3D printing requisite camera mounts, open source tools are available.

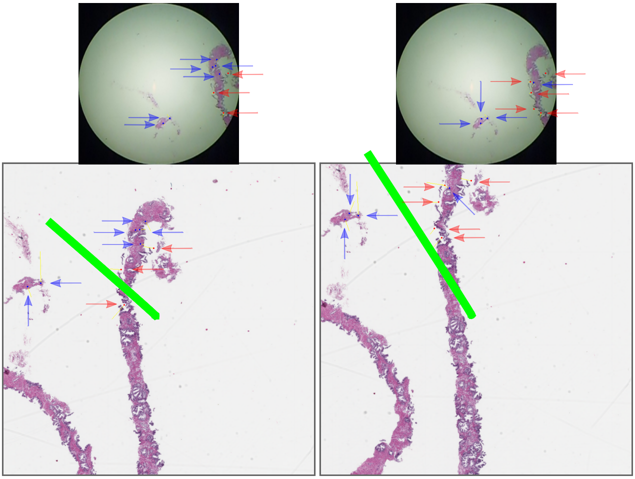


Figure 2.4: **Image registration.** The best image registration for a given video frame (same frame top left and top right) from the commodity camera at the microscope eyepiece compared to two different high-quality patches (bottom left and bottom right) from the whole slide scan image minimizes the length of the green line, which is the distance from the center of the patch to the center of the frame mapped into the patch's coordinate space. The green line's length is distance $d$ in Alg 2.1.

### 2.2.6 Video preprocessing and registration

A Debian Linux computer converted individual slide inspection video frames to PNG files using the ffmpeg program. OpenCV software detected slide movement via a dense optical flow procedure[43,44], comparing the current and preceding video frames, shown in Fig 2.3. Through this dense optical flow procedure. we calculated a movement vector for each pixel of each camera video frame, where a movement vector magnitude of one means the pixel has been displaced by one pixel in the video frame of interest, with respect to the previous video frame. Though the details of this procedure are beyond the scope of this work, a computationally efficient polynomial expansion method explains

```
input : $I_{frame}$: image from commodity camera, a video frame
        $I_{0,1,...,N-1}$: N overlapping patch images, together spanning whole slide
        $I_{prior} \in I_{0,1,...,N-1}$, the best matching patch from previous video frame
output: $I_{best} \in I_{0,1,...,N-1}$, the best matching patch to $I_{frame}$
$S_{frame} \longleftarrow$ set of all SURF interest points in $I_{frame}$;
$n \longleftarrow 0$, a counter through $I_{0,1,...,N-1}$ images;
$n_{best} \longleftarrow -1$, the value of $n$ where $I_n$ is $I_{best}$;
$d_{best} \longleftarrow MAXINT$, to store the distance between $I_{best}$ and $I_{frame}$ centers;
while $n < N$ do
    if $I_n$ is three or fewer patches spatially removed from $I_{prior}$ then
        $S_t \longleftarrow$ set of all SURF interest points in $I_n \in I_{0,1,...,N-1}$;
        $S_{fs} \longleftarrow$ subset of $S_{frame}$ points that match SURF feature vector of an $S_t$ point;
        $S_{ts} \longleftarrow$ subset of $S_t$ points that match SURF feature vector of an $S_{frame}$ point;
        $T \longleftarrow$ rigid body transformation of $I_{frame}$ pixel coordinate space into $I_n$ pixel coordinate space, calculated by point set registration of $RANSAC(S_{fs}, S_{ts})$;
        $d \longleftarrow$ distance in pixels between $I_{frame}$ center and $T(I_{frame})$ center, which measures how far off-center $I_n$ is from $I_{frame}$;
        if $d < d_{best}$ then
            $n_{best} \longleftarrow n$;
            $d_{best} \longleftarrow d$;
        end
    end
    $n \longleftarrow n + 1$;
end
return $I_{n \longleftarrow n_{best}}$, which is $I_{best}$;
```

**Algorithm 2.1: Automated image registration procedure (Fig 2.4) to find the least off-center patch from a given commodity camera video frame.** The whole slide image is split into N overlapping 800 × 800 px patches. "Three or fewer patches spatially removed" means any $I_n$ must be (i) $I_{prior}$, (ii) adjacent to $I_{prior}$, (iii) adjacent to a patch adjacent to $I_{prior}$, or (iv) adjacent to a patch adjacent to an $I_{prior}$-adjacent patch. In this way, $I_n$ is restricted to a spatial neighborhood localized around the prior match, typically improving image registration performance because most slide movements are small. On lens change, (i) the patch at lower magnification and (ii) the patches at higher magnification covering the same area as the current magnification's neighborhood are considered for registration only.

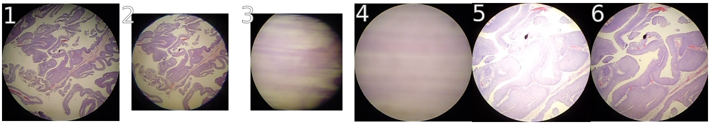


Figure 2.5: **Lens change detection.** The normal non-black pixel bounding box is initially 415 × 415 px. A change to 415 × 282 px indicates the pathologist changing the lens, thus changing slide magnification. Note some pixels that may appear black are called non-black due to difficult to perceive noise in the image, which effects calculated bounding box size. All images shown at same scale trimmed to bounding box.

a pixel's movement vector as the previous frame's pixel neighborhood polynomial transformed under translation to the current frame's pixel neighborhood polynomial, where a 39 × 39 px Gaussian weighting function averages pixel movement vectors for smoothing[43,44]. We defined slide movement to start if 10% or more of pixels in the entire field of view of the camera have a movement vector magnitude of at least one, and defined slide movement to stop if 2% or fewer of the pixels in the entire field of view of the camera have a movement magnitude vector of at least one. The entire field of view of the camera is 640 × 480 px, and a small subset of these capture the circular field of view at the microscope eyepiece, with the remaining pixels being black (Fig 2.3). The representative frame among consecutive unmoving frames moved the least. The ImageJ[45] SURF[46][1] and OpenCV software libraries registered each representative to an 800 × 800 px image patch taken from the high-resolution Aperio slide scanner. Each patch aggregated total pathologist time.

The partially automated registration process starts with initial manual registration of a frame, followed by automated registration within the preceding registration's spatial neighborhood (Fig 2.4 and Alg 2.1). First, (i) a set $S_{frame}$ of SURF interest points were found in the video frame, (ii) a set $S_t$ of SURF interest points were found in a slide image patch, (iii) SURF interest point feature vectors were compared in $S_{frame}$ and $S_t$ to determine which points were shared in $S_{frame}$ and $S_t$, and (iv) subsets of $S_{frame}$ and $S_t$ points that were shared were then stored in $S_{fs}$ and $S_{ts}$, respectively. Points shared between a camera video frame and an image patch (Fig 2.4 at left, top and bottom) change depending on the image patch (Fig 2.4 at right, top and bottom). Second, we used the OpenCV implementation of random sample consensus [RANSAC][47] for point set registration, to calculate a rigid body transformation from $S_{fs}$ point pixel positions in the video frame to $S_{ts}$ point pixel positions in the image patch, to find the distance in pixels that the video frame is off-center from the patch. Following this procedure for every image patch in the spatial neighborhood of the previous image registration, we selected the least off-center image patch as the best registration, because the pathologist's fovea is in approximately the same place in this video frame and image patch. Finally, a manual curation of registrations ensures correctness. Because slide movements are usually slight, this automated process reduces manual curation effort because automatic registrations are rarely far from the correct registration, so after the registration is corrected within a small localized neighborhood, automatic registrations may proceed from there. Fully automated image registration is not part of this study.

---

[1] ImageJ SURF is released under the GNU GPL and is available for download from http://labun.com/imagej-surf/

Slide magnification may change during inspection as the pathologist changes objective lenses. Lens change is detected automatically when the field of view bounding box of nonblack pixels changes size (Fig 2.5). SURF is scale-invariant so registrations may otherwise proceed at an unchanged magnification.

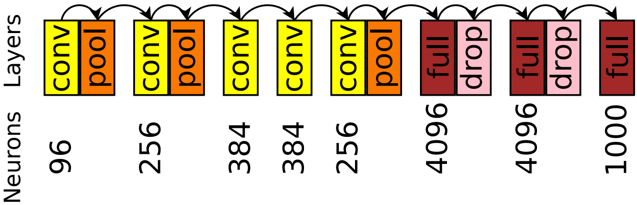


Figure 2.6: **CaffeNet architecture:** neuron counts, convolutional layers, pooling layers, dropout[48] layers, and fully-connected layers shown.

### 2.2.7 Deep learning

We used Caffe[49] for deep learning of convolutional features in a binary classification model given the 800 × 800 px image patches labeled with pathologist viewing times in seconds. To adapt for our purpose CaffeNet (Fig 2.6), which is similar to AlexNet[50], we re-initialized its top layer's weights after ImageNet[51] pre-training. Two output neurons were connected to the re-initialized layer, then training followed on augmented 800 × 800 px patches for 10,000 iterations in Caffe. In bladder, our model simply predicted whether or not a pathologist viewed an 800 × 800 px patch more than 0.1 s (30 fps camera). In prostate, due to the higher overlap between adjacent patches and less tissue available, to be salient a patch met at least one of these criteria: (1) viewed more than 0.1 s, (2) immediately above, below, left, or right of at least two patches viewed more than 0.1 s, or (3) above, below, left, right, or diagonal from at least three patches viewed more than 0.1 s such that all three are not on the same side. In this way, image patches highly overlapping in the neighborhood (Alg 2.1) of salient patches were not themselves considered nonsalient if a pathologist happened to jump over them during observation.

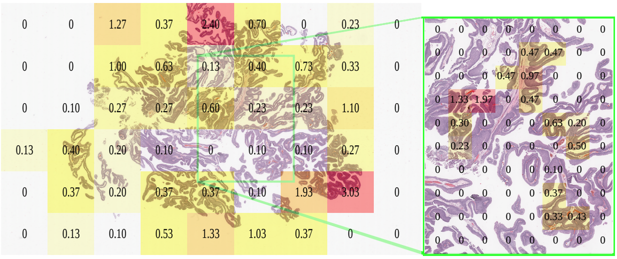

0 0 1.27 0.37 2.40 0.70 0 0.23 0
0 0 1.00 0.63 0.13 0.40 0.73 0.33 0
0 0.10 0.27 0.27 0.60 0.23 0.23 1.10 0
0.13 0.40 0.20 0.10 0 0.10 0.10 0.27 0
0 0.37 0.20 0.37 0.37 0.10 1.93 3.03 0
0 0.13 0.10 0.53 1.33 1.03 0.37 0 0
0 0 0 0 0 0 0 0
0 0 0 0 0.47 0.47 0 0
0 0 0 0.47 0.97 0 0 0
0 1.33 1.97 0 0.47 0 0 0
0 0.30 0 0 0 0.63 0.20 0
0 0.23 0 0 0 0 0.50 0
0 0 0 0 0 0.10 0 0
0 0 0 0 0 0.37 0 0
0 0 0 0 0 0.33 0.43 0
0 0 0 0 0 0 0 0

Figure 2.7: **Pathologist viewing times in seconds at the microscope** for low (*left*, 10x, level 2) and high magnification (*right*, 20x, level 1), registered to the same urothelial carcinoma slide scan.

### 2.2.8 Slide magnifications and spatial partitions

Urothelial carcinoma (bladder) in Fig 2.7 was analyzed first, with author HAA inspecting at the microscope. Viewed regions at the microscope corresponded to the whole-slide scan SVS file at magnification levels 2 and 1. We restricted our analysis to level 2, having insufficient level 1 data. We split level 2 into three portions: left, center, and right. Due to over 50% overlap among the slide's total 54 800 × 800 px level 2 patches, we excluded the center portion from analysis, but retained left and right sides, which did not overlap (Fig 2.8).

### 2.2.9 Viewing time threshold for positive/negative ground truth

In bladder, we considered a negative example to be a patch viewed for 0.1 s (3 frames or fewer, 30 fps) or less, and a positive example viewed for more than 0.1 s (4 frames or more). This produced 9 positive and 9 negative examples on the left side, and the same number on the right. We performed three-fold

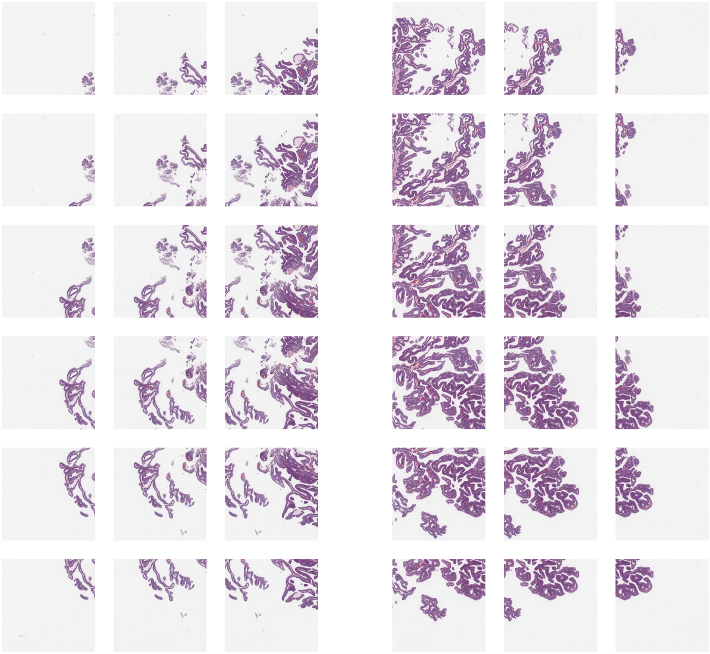


Figure 2.8: **Scaled image patches of left and right sides of bladder patient 1 slide (Fig 2.2).** Middle excluded here and not used in analysis, to isolate left and right sides from each other. Note far left and far right have less tissue, but tissue is present for training. The overlap among patches is evenly distributed and greater than 50%.

cross-validation on the left (6+ and 6- examples training set, 3+ and 3- examples validation set), then used the model with the highest validation accuracy on the right to calculate test accuracy, an estimate of generalization error (Fig 2.9). This cross-validation was duplicated ten times on the left, each time estimating test accuracy, to calculate a confidence interval. We then duplicated this training/validating on the left and testing on the right.

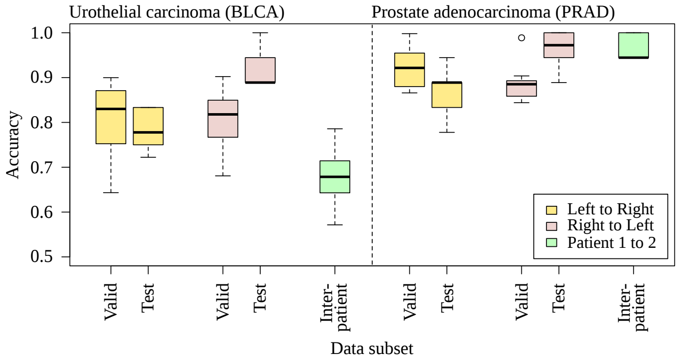


Figure 2.9: **Ten three-fold cross-validation trials for bladder [BLCA] and prostate [PRAD]**, evaluated for intrapatient training/validating on left while testing on the right and vice versa. Each model is evaluated against a different patient (interpatient), slides in Fig 2.2). The needle for prostate cancer biopsy may standardize the distribution of prostate tissue in the whole slide, maintaining a higher accuracy of the prostate classifier on an interpatient basis than the bladder cancer classifier. The bladder patients are transurethral resections taken by cuts rather than a standard gauge needle.

### 2.2.10 Data augmentation

Training and validation data were augmented. For a 800 × 800 px patch, all flips and one-degree rotations through 360 degrees were saved, then cropped to the centermost 512 × 512 px, then scaled to 256 × 256 px. This rotation-based data augmentation biases the neural network to learn rotationally-invariant features rather than overfit to the training data's particular orientation, e.g. the angle of prostate needle biopsy tissue strips. Thus intrapatient and interpatient test sets are not augmented, but training and validation sets are augmented. The cancer diagnosis or viewing time in pathology is not expected to change when rotating or flipping a slide. We direct readers to Krizhevsky *et al.* 2012[50] for more information on data augmentation. Like Krizhevsky's data augmentation of 224 × 224 px random crops for small translations, we further augment our dataset through random crops of 227×227 px, which is the default for CaffeNet. Unlike Krizhevsky, we do not augment our dataset through minor perturbations in the principal components of the RGB color space.

In bladder, the augmented training set size was 8,640 patches. This 8,640 count includes rotations and flips, but does not include random crops, which were performed automatically by Caffe at training time. Caffe randomly cropped 256×256 px patches to 227×227 px for each iteration of CaffeNet learning. No images in the validation set were derived from the training set, and vice versa. A training set consists of two concatenated folds, with the remaining fold as validation. In addition to the bladder cancer slide, we analogously processed two prostate cancer needle biopsy slides, with author SJS inspecting these slides. In prostate, the augmented training set size was 8,160 patches.

### 2.2.11 Cross-validation and testing

Training and validation sets are drawn from the same side of the slide, i.e. both sets on the left or both sets on the right (Fig 2.8). Patches in a training set may have at least 50% overlap with patches in a validation set. Overlapping regions of these images have identical sets of pixels, guaranteeing training and validation sets are exchangeable for valid cross-validation. If training error steadily decreases while validation error steadily increases, where training and validation sets are exchangeable, then the classifier is overfit. In contrast, the other side of the slide is used as a test set and may appear obviously different than the training and validation sets, e.g. the left side of Fig 2.8 appears different than the right side. We test the other side to estimate generalization error, which measures how the classifier may perform on data unseen at training time. Testing on the other side of the slide guarantees there is no overlap with the training set, so the test data is unseen by the classifier at training time.

Different cross-validation schemes are conceivable, such as (i) a top versus bottom split rather than a left versus right split or (ii) a leave-one-out [LOO] cross-validation approach. Unfortunately, Fig 2.8 shows a slight overlap in the row second from the top and the row second from the bottom, effectively reducing by 25-50% the amount of data for training, validation, and testing compared to our left versus right approach. Separately, in a LOO setting, one may draw a test patch, then draw training and validation sets randomly that do not overlap with the test patch, keeping training and validation set sizes constant for every possible test patch in the slide. Unfavorably, if the test patch is drawn from the middle column of the slide, then only the leftmost and rightmost columns of patches do not overlap with the test patch, reducing the amount of data for training and validation sets by 33% compared to our left versus right approach. This 33% reduction for middle test patches is in contrast to the 111% increase in training and validation data quantity for test patches drawn from the corners of the leftmost or rightmost columns, where this excess is randomly discarded to maintain constant training and validation set sizes for all possible test patches. Moreover, if the test patch is in the bottom row on the right side, the top row on the right side may be sampled for training, which may inflate the LOO generalization accuracy estimate compared to our cross-validation approach that trains only on the left when testing on the right, due to patches on the right appearing similar to one another. We show in Section 2.3 that training on the left and testing on the right gives significantly different accuracy compared to training on the right and testing on the left, suggesting the left and right sides have indeed

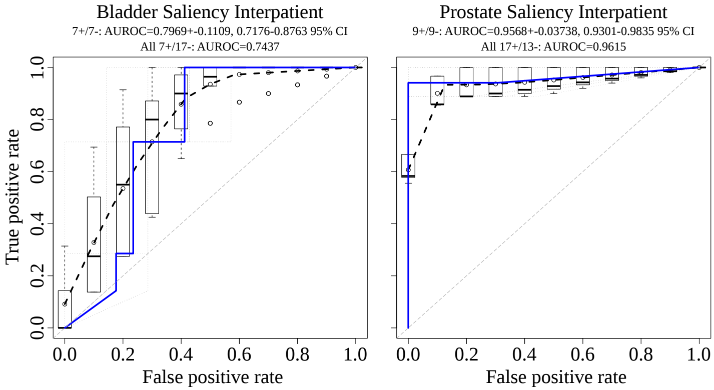


Figure 2.10: **Interpatient area under the receiver operating characteristic [AUROC] for bladder and prostate**, with dashed black curve for average AUROC over draws of the data and blue line for all data used from the patient.

different distributions of information. Thus compared to these alternatives, our left versus right three-fold cross-validation approach (i) maximizes the sizes of the training, validation, and test sets, (ii) conservatively estimates generalization error by not training the classifier on data that appear similar to the test set, and (iii) samples each patch on the left or right sides exactly once for an overall validation error measure for that side.

## 2.3 Results

In bladder, when training/validating on the left side and testing on the right, mean test accuracy is 0.781±0.0423 (stdev) with 95% confidence interval [CI] from 0.750 to 0.811 (df=9, Student's T, Table 2.1). When training/validating on the right and testing on the left, mean test accuracy is 0.922±0.0468 with 0.889-0.956 95% CI (Table 2.1). Overall mean test accuracy is 85.15%. The left and right

Table 2.1: **Accuracies of ten trials of three-fold cross-validation in bladder.** Validation and test accuracies for a single slide video of urothelial carcinoma (patient 1, slide at upper left in Fig 2.2, performance plotted at left in Fig 2.9)), left side of the slide versus right side.

Column “Fold0 Valid” reports validation accuracy when folds 1 and 2 were used for training. Similarly, “Fold1 Valid” is for folds 0 and 2 training. “Valid Acc” is the validation accuracy overall – the average of “Fold0 Valid”, “Fold1 Valid”, and “Fold2 Valid”. Because we will use a single classifier for saliency prediction, we selected the classifier with highest validation accuracy and highlighted it yellow, e.g. we selected the Fold0 classifier with 0.9218 validation error as shown in the third row, namely Trial 2 leftright.

Column “Fold0 Test” reports the test accuracy of the classifier trained on folds 1 and 2. Because we will use a single classifier not an ensemble, we highlight the test accuracy of the classifier selected by highest validation accuracy and report this in “Test Acc“ as generalization accuracy, e.g. we selected the Trial 2 leftright Fold0 classifier having “Fold0 Test” of 0.8333 and copied this to “Test Acc”. We report test accuracies for all three classifiers, showing Fold1 and Fold2 classifiers tie for highest validation accuracy in Trial 1 leftright, so their test accuracies of 0.7222 and 0.7778 were averaged for a “Test Acc” of 0.7500. As another sanity check in our small data setting, we report that the variance in the selected-versus-non-selected test accuracy differences is not greater than the selected-versus-non-selected validation accuracy differences (F-Test p=0.5662 and Bartlett’s Test p=0.5661. Validation differences normally distributed by Anderson-Darling p=0.08837, and test differences by p=0.1734). If it were greater, there may be experimental setup problems because training would not be stably producing classifiers that learn the saliency concept. Finally, one may train a classifier on all folds then evaluate test accuracy with this classifier, but a performance boost from additional training data may inflate generalization accuracy. In Sec 2.3 we show without such inflation there remains a significant difference in generalization accuracy and interpatient accuracy in bladder.

Testing the best classifier (highlighted in cyan, highest test accuracy on this and other folds, secondarily highest mean validation accuracy) on draws of the data on the second bladder patient, accuracies are 0.643, 0.786, 0.714, 0.786, 0.714, 0.714, 0.643, 0.571, 0.643, and 0.571.

| Direction | Trial | Fold0 Valid | Fold1 Valid | Fold2 Valid | Valid Acc | Fold0 Test | Fold1 Test | Fold2 Test | Test Acc |
|---|---|---|---|---|---|---|---|---|---|
| leftright | 0 | 0.9466 | 0.5850 | 0.9680 | 0.8332 | 0.8333 | 0.7222 | 0.7778 | 0.7778 |
| leftright | 1 | 0.8852 | 0.9070 | 0.9070 | 0.8997 | 0.7778 | 0.7222 | 0.7778 | 0.7500 |
| leftright | 2 | 0.9218 | 0.8602 | 0.8832 | 0.8884 | 0.8333 | 0.7222 | 0.7778 | 0.8333 |
| leftright | 3 | 0.7640 | 0.7812 | 0.7120 | 0.7524 | 0.7778 | 0.7778 | 0.7778 | 0.7778 |
| leftright | 4 | 0.7590 | 0.6576 | 0.5134 | 0.6433 | 0.8333 | 0.7778 | 0.7778 | 0.8333 |
| leftright | 5 | 0.9268 | 0.7416 | 0.5088 | 0.7257 | 0.8333 | 0.7778 | 0.7778 | 0.8333 |
| leftright | 6 | 0.8028 | 0.7988 | 0.7048 | 0.7688 | 0.7222 | 0.7778 | 0.7778 | 0.7222 |
| leftright | 7 | 0.7318 | 0.8402 | 0.9088 | 0.8269 | 0.7778 | 0.7778 | 0.7778 | 0.7778 |
| leftright | 8 | 0.9572 | 0.7608 | 0.8418 | 0.8533 | 0.7778 | 0.8333 | 0.8333 | 0.7778 |
| leftright | 9 | 0.7492 | 0.8774 | 0.9860 | 0.8709 | 0.7778 | 0.7778 | 0.7222 | 0.7222 |
| rightleft | 0 | 0.8802 | 0.8528 | 0.8554 | 0.8628 | 1.0000 | 0.8889 | 0.9444 | 1.0000 |
| rightleft | 1 | 0.7662 | 0.5982 | 0.9364 | 0.7669 | 0.8889 | 0.9444 | 1.0000 | 1.0000 |
| rightleft | 2 | 0.9492 | 0.8560 | 0.7308 | 0.8453 | 0.9444 | 0.8889 | 0.9444 | 0.9444 |
| rightleft | 3 | 0.5404 | 0.8206 | 0.8368 | 0.7326 | 0.9444 | 0.9444 | 0.8889 | 0.8889 |
| rightleft | 4 | 0.6560 | 0.7114 | 0.6748 | 0.6807 | 0.8889 | 0.8889 | 0.8333 | 0.8889 |
| rightleft | 5 | 0.8932 | 0.7062 | 0.7310 | 0.7768 | 0.9444 | 0.8889 | 0.8333 | 0.9444 |
| rightleft | 6 | 0.8560 | 0.8540 | 0.9966 | 0.9022 | 0.8889 | 0.9444 | 0.8889 | 0.8889 |
| rightleft | 7 | 0.8362 | 0.8560 | 0.7978 | 0.8300 | 0.8333 | 0.8889 | 1.0000 | 0.8889 |
| rightleft | 8 | 0.7200 | 0.8546 | 0.9740 | 0.8495 | 1.0000 | 0.9444 | 0.8889 | 0.8889 |
| rightleft | 9 | 0.8634 | 0.8634 | 0.6904 | 0.8057 | 0.8333 | 0.9444 | 1.0000 | 0.8889 |

Table 2.2: **Accuracies of ten trials of three-fold cross-validation in prostate.** Validation and test accuracies for a single slide video of prostate adenocarcinoma (patient 1, slide at upper right in Fig 2.2, performance plotted at right in Fig 2.9)), left side of the slide versus right side.

Testing the best classifier on draws of the data on the second prostate patient, accuracies are 0.944, 1, 0.944, 0.944, 1, 0.944, 0.944, 0.944, 1, and 1.

| Direction | Trial | Fold0 Valid | Fold1 Valid | Fold2 Valid | Valid Acc | Fold0 Test | Fold1 Test | Fold2 Test | Test Acc |
|---|---|---|---|---|---|---|---|---|---|
| leftright | 0 | 0.9992 | 0.9946 | 1.0000 | 0.9979 | 0.7778 | 0.7778 | 0.7778 | 0.7778 |
| leftright | 1 | 0.9512 | 0.7282 | 0.9994 | 0.8929 | 0.8889 | 0.8889 | 0.8889 | 0.8889 |
| leftright | 2 | 0.9550 | 1.0000 | 0.6530 | 0.8693 | 0.8889 | 0.7778 | 0.7222 | 0.7778 |
| leftright | 3 | 0.8636 | 0.9992 | 1.0000 | 0.9543 | 0.8889 | 0.7778 | 0.9444 | 0.9444 |
| leftright | 4 | 0.8276 | 0.7760 | 0.9940 | 0.8659 | 0.8889 | 0.7778 | 0.8889 | 0.8889 |
| leftright | 5 | 0.8654 | 0.9986 | 1.0000 | 0.9547 | 0.9444 | 0.9444 | 0.9444 | 0.9444 |
| leftright | 6 | 0.8560 | 0.9862 | 0.9992 | 0.9471 | 0.8889 | 0.8889 | 0.8889 | 0.8889 |
| leftright | 7 | 0.8674 | 1.0000 | 0.9984 | 0.9553 | 0.8889 | 0.8333 | 0.8889 | 0.8333 |
| leftright | 8 | 0.9560 | 0.8560 | 0.8760 | 0.8960 | 0.8889 | 0.8333 | 0.9444 | 0.8889 |
| leftright | 9 | 0.6846 | 0.9992 | 0.9560 | 0.8799 | 0.8333 | 0.8333 | 1.0000 | 0.8333 |
| rightleft | 0 | 0.9786 | 0.7760 | 0.9146 | 0.8897 | 0.8889 | 0.9444 | 1.0000 | 0.8889 |
| rightleft | 1 | 1.0000 | 0.7292 | 0.8460 | 0.8584 | 1.0000 | 0.9444 | 1.0000 | 1.0000 |
| rightleft | 2 | 0.7130 | 0.9512 | 0.8676 | 0.8439 | 0.8889 | 1.0000 | 0.8333 | 1.0000 |
| rightleft | 3 | 0.9998 | 1.0000 | 0.9664 | 0.9887 | 0.9444 | 0.9444 | 1.0000 | 0.9444 |
| rightleft | 4 | 0.7760 | 1.0000 | 0.8842 | 0.8867 | 0.8889 | 0.9444 | 1.0000 | 0.9444 |
| rightleft | 5 | 0.9758 | 0.9984 | 0.5926 | 0.8556 | 0.9444 | 0.8889 | 0.9444 | 0.8889 |
| rightleft | 6 | 0.6344 | 0.9770 | 1.0000 | 0.8705 | 0.8889 | 1.0000 | 1.0000 | 1.0000 |
| rightleft | 7 | 0.7760 | 0.9028 | 1.0000 | 0.8929 | 0.8889 | 1.0000 | 1.0000 | 1.0000 |
| rightleft | 8 | 0.8560 | 0.8560 | 0.9992 | 0.9037 | 0.9444 | 0.9444 | 0.9444 | 0.9444 |
| rightleft | 9 | 0.8560 | 0.9412 | 0.8538 | 0.8837 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |

test accuracies differ (p=0.000135, Wilcoxon rank-sum, n=20), while validation accuracies do not (p=0.9118, n=20). This suggests nonhomogenous information content throughout the slide. Indeed, the pathologist started and ended slide inspection on the right, and spent double the time on the right versus the left (Fig 2.7, 8.32 s right, 4.07 s left). The second bladder had different morphology and model accuracy reduced to 0.678±0.0772, 0.623-0.734 95% CI. Moreover, the second bladder had only 7 positive examples available, whereas both prostates and the first bladder had at least 9 positive examples.

For the first prostate slide, training on the left side and testing on the right, we find accuracy 0.867±0.0597, 0.824-0.909 95% CI (Table 2.2). Training on the right and testing on left, we find 0.961±0.0457, 0.928-0.994 95% CI (Table 2.2). Overall mean test accuracy is 91.40%. Taking the best model learned from this first prostate (right side, test accuracy 100%, 18/18), we tested on the second prostates right side (because the left did not have 9 positive training examples) and find 0.967±0.0287, 0.946-0.987 95% CI. We also tested this model on the bladder cancer slide, and find 0.780 accuracy on the left and 0.720 on the right (9+ and 9- training examples each), mean accuracy 75.00%. The best bladder cancer model predicts every patch is not salient in both prostates, presumably because the little tissue in prostate is insufficient for a positive saliency prediction.

Interpatient AUROC for bladder and prostate is shown in Fig 2.10. In prostate, nine salient and nine nonsalient examples are drawn from the second patient. Average AUROC was calculated from ten such draws, achieving a mean±stdev of 0.9568±0.0374 and 95% CI of 0.9301-0.9835. Over all 17 salient and 13 nonsalient patches used from the second prostate patient, the AUROC is 0.9615. In bladder, due to fewer patches available in the small slide, only seven salient and seven nonsalient examples are drawn from the second patient. Average AUROC was calculated for ten such draws, achieving 0.7929±0.1109 and 95% CI of 0.7176-0.8763. Over all 7 salient and 17 nonsalient patches used from the second bladder patient, the AUROC is 0.7437. These nonoverlapping confidence intervals are evidence the bladder cancer classifier distinguishes salient from nonsalient patches less well than the prostate cancer classifier, and a Wilcoxon rank-sum test indeed finds the difference in classifier performance by these ten draws each from bladder and prostate is significant (p=0.0001325) (Fig 2.9).

The deep convolutional network CaffeNet emits a score from 0 to 1 when predicting if an image patch is salient or not. When taking a score of greater than 0.5 to be salient, the p-value from Fisher's Exact Test is 1.167e-7 in prostate (16 true positives, 1 false negative, 0 false positives, 13 true negatives) and 0.009916 in bladder (7 true positives, 0 false negatives, 7 false positives, 10 true negatives), indicating our trained CaffeNet classifier in both tissues accurately distinguishes salient from nonsalient regions when trained on one patient and predicting in another.

## 2.4 Conclusion

Collecting image-based expert annotations for the deluge of medical data at modern hospitals is one of the tightest bottlenecks for the application of large-scale supervised machine learning. We address this with a novel framework that combines a commodity camera, 3D-printed mount, and software stack to build a predictive model for saliency on whole slides, i.e. where a pathologist looks to make a diagnosis. The registered regions from the digital slide scan are markedly higher quality than the camera frames, since they do not suffer from debris, vignetting, and other artifacts. The proposed CNN is able to predict salient slide regions with a test accuracy of 85-91%. We plan to scale up this pilot study to more patients, tissues, and pathologists.

## Acknowledgments

AJS was supported by NIH/NCI grant F31CA214029 and the Tri-Institutional Training Program in Computational Biology and Medicine (via NIH training grant T32GM083937). This research was funded in part through the NIH/NCI Cancer Center Support Grant P30CA008748. AJS thanks Terrie Wheeler, Du Cheng, and the Medical Student Executive Committee of Weill Cornell Medical College for free 3D printing access, instruction, and support. AJS thanks Mariam Aly for taking the photo of the camera on the orange 3D-printed mount in Fig 2.1, and attention discussion. We acknowledge fair use of part of a doctor stick figure image in Fig 2.1 from 123rf.com. AJS thanks Mark Rubin for helpful pathology discussion. AJS thanks Paul Tatarsky and Juan Perin for Caffe install help on the Memorial Sloan Kettering supercomputer. We gratefully acknowledge NVIDIA Corporation for providing us a GPU as part of the GPU Research Center award to TJF, and for their support with other GPUs.

# CHAPTER 3

# SPOP MUTATION PREDICTION FROM PROSTATE CANCER HISTOPATHOLOGY

Portions of this chapter first appeared in Schaumberg* et al 2016. Andrew J. Schaumberg originally wrote Schaumberg et al 2016, which was editted by Thomas J. Fuchs and further reviewed by Mark A. Rubin[1]. The author makes this chapter available under the terms of a Creative Commons Attribution-NonCommercial-NoDerivatives [CC-BY-NC-ND] 4.0 International license, at `https://creativecommons.org/licenses/by-nc-nd/4.0/`.

## 3.1 Introduction

Genetic drivers of cancer morphology, such as E-Cadherin [CDH1] loss promoting lobular rather than ductal phenotypes in breast, are well known. TMPRSS2-ERG fusion in prostate cancer has a number of known morphological traits, including blue-tinged mucin, cribriform pattern, and macronuclei[54]. Computational pathology methods[25] typically predict clinical or genetic features as a function of histological imagery, e.g. whole slide images. Our central hypothesis is that the morphology shown in these whole slide images, having nothing more than standard hematoxylin and eosin [H&E] staining, is a function of the underlying genetic drivers. To test this hypothesis, we gathered a cohort of 499 prostate adenocarcinoma patients from The Cancer Genome Atlas [TCGA][2], 177 of which were suitable for analysis, with 20 of those having mutant SPOP (Figs 3.1, 3.2, and S3.7). We then used ensembles of deep convolutional neural

---

[*] Schaumberg A, Rubin M, and Fuchs T. H&E-stained Whole Slide Deep Learning Predicts SPOP Mutation State in Prostate Cancer. *bioRxiv*, page 064279, July 2016. doi:10.1101/064279.

[1] The ongoing and developing nature of this work is in part reflected in the slight title changes to the preprint, which is available at `https://doi.org/10.1101/064279`

[2] TCGA data courtesy the TCGA Research Network http://cancergenome.nih.gov/

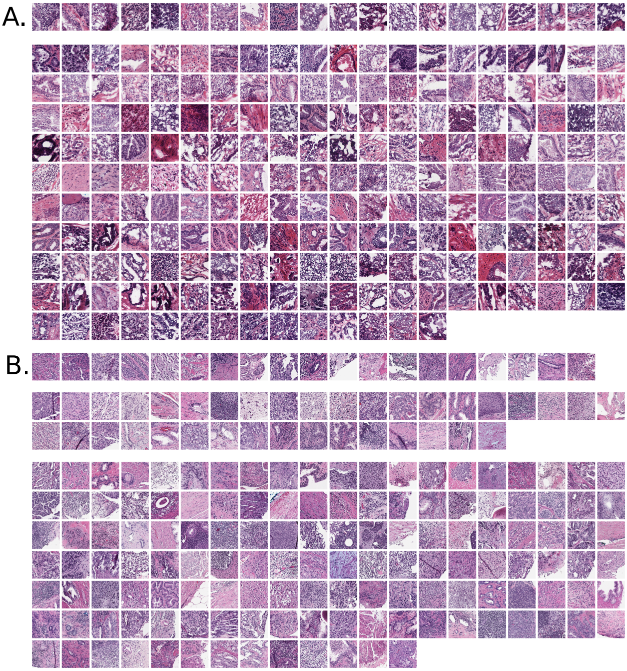


Figure 3.1: **TCGA and MSK-IMPACT slides for SPOP study. (A)** TCGA cohort of frozen section images. Top row shows 20 SPOP mutants. Bottom rows are 157 SPOP non-mutants, where 25 patients had 2 and 6 patients had 3 acceptable slides available. **(B)** MSK-IMPACT cohort of formalin-fixed paraffin-embedded sections, providing higher image quality than frozens. Top row shows 19 SPOP mutants. Middle rows show 36 SPOP mutants scanned as added training data for TCGA testing. Bottom rows are 133 SPOP non-mutants.

networks to accurately predict whether or not SPOP was mutated in the patient, given only the patient's whole slide image (Figs  and 3.4A), leveraging spatial localization of SPOP mutation evidence in the histology imagery (Fig 3.4B,C) for statistically significant SPOP mutation prediction accuracy when training on TCGA but testing on the MSK-IMPACT[55] cohort (Fig 3.5). Further, we scanned 36 additional SPOP mutant MSK-IMPACT slides, training on this ex-

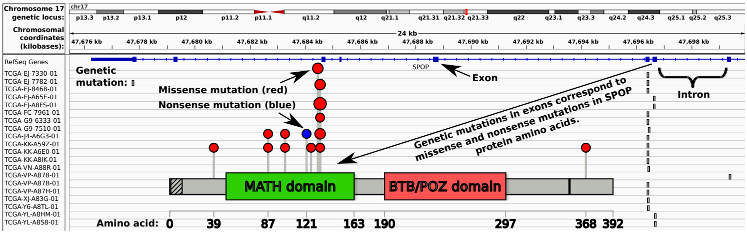

Chromosome 17 genetic locus:
chr17
p13.3
p13.2
p13.1
p12
p11.2
p11.1
q11.2
q12
q21.1
q21.31
q21.32
q21.33
q22
q23.1
q23.3
q24.2
q24.3
q25.1
q25.2
q25.3
Chromosomal coordinates (kilobases):
24 kb
47,676 kb
47,678 kb
47,680 kb
47,682 kb
47,684 kb
47,686 kb
47,688 kb
47,690 kb
47,692 kb
47,694 kb
47,696 kb
47,698 kb
RefSeq Genes
SPOP
Exon
Intron
TCGA-EJ-7330-01
TCGA-EJ-7782-01
TCGA-EJ-8468-01
TCGA-EJ-A65E-01
TCGA-EJ-A8FS-01
TCGA-FC-7961-01
TCGA-G9-6333-01
TCGA-G9-7510-01
TCGA-J4-A6G3-01
TCGA-KK-A59Z-01
TCGA-KK-A6E0-01
TCGA-KK-A8IK-01
TCGA-VN-A88R-01
TCGA-VP-A878-01
TCGA-VP-A87B-01
TCGA-VP-A87H-01
TCGA-XJ-A83G-01
TCGA-Y6-A8TL-01
TCGA-YL-A8HM-01
TCGA-YL-A8S8-01
Genetic mutation: ▯
Missense mutation (red)
Nonsense mutation (blue)
Genetic mutations in exons correspond to missense and nonsense mutations in SPOP protein amino acids.
MATH domain
BTB/POZ domain
Amino acid: 0 39 87 121 163 190 297 368 392

Figure 3.2: **SPOP mutations in the Integrated Genomics Viewer**[52,53], with lollipop plot showing mutations in most frequently mutated domain. For two of twenty patients, somatic SPOP mutations fall outside the MATH domain, responsible for recruiting substrates for ubiquitinylation.

panded MSK-IMPACT cohort and testing on the TCGA cohort. Our classifier's generalization error bounds (Fig 3.5A,B), receiver operating characteristic (Fig 3.5C1,D1), and independent dataset performance (Fig 3.5C2,D2) support our hypothesis, in agreement with earlier work suggesting SPOP mutants are a distinct subtype of prostate cancer[56]. Finally, we applied our metaensemble classifier to the content-based image retrieval [CBIR] task of finding similar patients to a given query patient (Fig 3.6), according to SPOP morphology features evident in the patient slide dominant tumor morphology.

Previously, pathologists described slide image morphologies, then correlated these to molecular aberrations, e.g. mutations and copy number alterations[57,58]. Our deep learning approach instead learns features from the images without a pathologist, using one mutation as a class label, and quantifies prediction uncertainty with confidence intervals [CIs] (Fig ).

Others used support vector machines to predict molecular subtypes in a bag-of-features approach over Gabor filters[59]. The authors avoided deep learning due to limited data available. Gabor filters resemble first layer features in a convolutional network. A main contribution of ours is using pre-training, Monte

Carlo cross-validation, and heterogeneous ensembles to enable deep learning despite limited data. We believe our method's prediction of a single mutation is more clinically actionable than predicting broad gene expression subtypes.

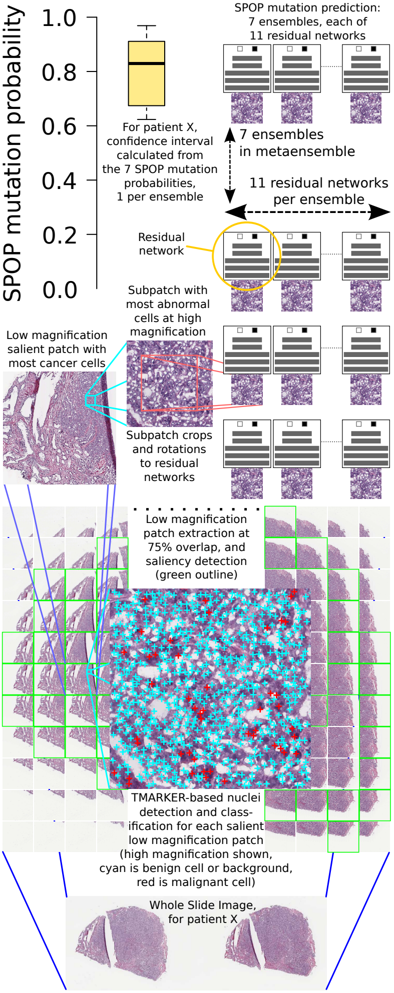

SPOP mutation probability
1.0
0.8
0.6
0.4
0.2
0.0
For patient X, confidence interval calculated from the 7 SPOP mutation probabilities, 1 per ensemble
SPOP mutation prediction: 7 ensembles, each of 11 residual networks
7 ensembles in metaensemble
11 residual networks per ensemble
Residual network
Subpatch with most abnormal cells at high magnification
Low magnification salient patch with most cancer cells
Subpatch crops and rotations to residual networks
Low magnification patch extraction at 75% overlap, and saliency detection (green outline)
TMARKER-based nuclei detection and classification for each salient low magnification patch (high magnification shown, cyan is benign cell or background, red is malignant cell)
Whole Slide Image, for patient X

Figure 3.3: **SPOP mutation prediction pipeline.** A whole slide image is split into patches (bottom) at low magnification. Salient patches are identified. The salient patch with the most cancer cells is deemed the "dominant tumor" patch and further analyzed. At high magnification, a sliding window within the dominant tumor patch finds the region with maximum abnormal cells. Deep neural networks then predict SPOP mutation and a confidence interval is calculated over these predictions.

Support vector machines, Gaussian mixture modeling, and principal component analyses have predicted PTEN deletion and copy number variation in cancer, but relied on fluorescence *in situ* hybridization [FISH], a very specific stain[60]. Our approach uses standard H&E, a non-specific stain that we believe could be utilized to predict more molecular aberrations than

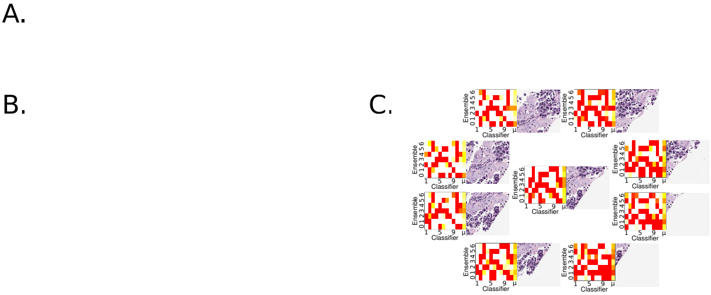


Figure 3.4: **Deep learning architecture and activation heatmaps.** **(A)** ResNet-50 architecture[62], top. Customized Drop50 architecture supports ResNet-50 pretraining, but has additional dropout[48] and fully connected neuron layers. In practice at least one of these architectures converges to a validation accuracy of 0.6 or more. Convolutional layers "conv", fully-connected layers "fc", 50% dropout layer "drop". ResNet-50 has 26,560 neurons and Drop50 has 28,574. All seven trials, each trial being a multiarchitectural ensemble of eleven residual networks, is 2,138,630 neurons total for TCGA training and MSK-IMPACT testing, and 2,132,858 neurons total for MSK-IMPACT training and TCGA testing. **(B)** Region of interest and surrounding octagon patches, each from an 800x800 pixel [px] patch cropped to 512x512px and scaled to 256x256px. At upper left, histological imagery leads to strong SPOP mutation predictions shown in red. At lower right, no such evidence exists and SPOP mutation is not predicted here, indicated in heatmaps as white rather than red. For each of the nine patches, the weighted mean prediction is calculated, shown in the heatmaps as the $\mu$ column at right. Each classifier in the ensemble makes a prediction for a patch, and classifiers having greater prediction variance among the nine patches are more weighted in the means for the patches. The metaensemble's SPOP mutation prediction is 0.6244, with 95% CI of 0.5218-0.7211 and 99% CI of 0.4949-0.7489, so this patient's tumor is predicted to have an SPOP mutation at 95% confidence but not 99% confidence. **(C)** Another MSK-IMPACT patient shown for the TCGA-trained metaensemble, suggesting there is greater SPOP mutation histological evidence in the lower right. The metaensemble's SPOP mutation prediction is 0.5528, with 95% CI of 0.5219-0.5867 and 99% CI of 0.5128-0.5962, so there is 99% confidence of SPOP mutation.

only the SPOP mutation that is our focus here. However, our method does not quantify tumor heterogeneity.

Tumor heterogeneity has been analyzed statistically by other groups[61], as a spatial analysis of cell type entropy cluster counts in H&E-stained whole slides. A high count, when combined with genetic mutations such as TP53, improves patient survival prediction. This count is shown to be independent of underlying genetic characteristics, whereas our method predicts a genetic characteristic, i.e. SPOP mutation, from convolutional features of the imaging.

Clustering patients according to hand-engineered features has been prior practice in histopathology CBIR, with multiple pathologists providing search relevancy annotations to tune the search algorithm[63]. Our approach relies on neither pathologists nor feature engineers, and instead learns discriminative genetic-histologic relationships in the dominant tumor to find similar patients. We also do not require a pathologist to identify the dominant tumor, so our CBIR search is automated on a whole slide basis. Because the entire slide is the query, we do not require human judgement to formulate a search query, so CBIR search results may be precalculated and stored for fast lookup.

## 3.2 Results

### 3.2.1 Molecular information as labels of pathology images opens a new field of molecular pathology

Rather than correlating or combining genetic and histologic data, we predict a gene mutation directly from a whole slide image with unbiased H&E stain. Our methods enable systematic investigation of other genotype and phenotype relationships, and serve as a new supervised learning paradigm for clinically actionable molecular targets, independent of clinician-supplied labels of the histology. Epigenetic, copy number alteration, gene expression, and post-

Figure 3.5: SPOP prediction performance. Stricter CIs reduce false negatives and maintain significant Fisher's Exact p-values, but ignore more predictions as inconclusive (panels (C2), (C3), (D2), (D3) – and Alg 3.3). The metaensemble (C1) consists of seven ensembles, with Receiver Operating Characteristics for each shown in Fig S3.9. The MSK-IMPACT training to TCGA testing performance ($0.01 < p < 0.05$, panels (D2) and (D3)) is expected due to MSK-IMPACT fixed sections being higher quality than TCGA frozen sections (Fig 3.1). Freezing distorts the slide by effecting watery and fatty tissues differently. Training on distorted data (TCGA) but testing on undistorted data (MSK-IMPACT) appears robust ($0.001 < p < 0.01$, panels (C2) and (C3)) as expected, due to the classifiers learning SPOP-discriminative features despite distortions in the data. This is similar in principle to de-noising autoencoders learning to be robust to noise in their data.

translational modification data may all label histology images for supervised learning. Future work may refine these predictions to single-cell resolution, combined with single-cell sequencing, immunohistochemistry, FISH, mass cytometry[64], or other technologies to label corresponding H&E images or regions therein. We suggest focusing on labels that are clinically actionable, such as gains or losses of function.

### 3.2.2 SPOP mutation state prediction is learnable from a small set of whole slides stained with hematoxylin and eosin

Despite SPOP being one of the most frequently mutated genes in prostate adenocarcinomas[56], from a TCGA cohort of 499 patients only 177 passed our quality control (Alg 3.1) and only 20 of these had SPOP mutation. Meanwhile in the 152-patient MSK-IMPACT cohort there were only 19 SPOP mutants, and though we could scan an additional 36 SPOP mutant archived slides, there are difficulties in practice acquiring large cohorts of patients with both quality whole slide cancer pathology images and genetic sequencing represented. This challenge increases for rare genetic variants. Moreover, different cohorts may have different slide preparations and appearances, such as TCGA being frozen sections and MSK-IMPACT being higher quality formalin-fixed paraffin embedded sections (Fig 3.1). Nonetheless, our pipeline (Fig ) accurately predicts whether or not SPOP is mutated in the MSK-IMPACT cohort when trained on the TCGA cohort (Fig 3.5), and vice versa. We leverage pre-trained neural networks, Monte Carlo cross-validation, class-balanced stratified sampling, and architecturally heterogeneous ensembles for deep learning in this “small data” setting of only

20 positive examples, which may remain relevant for rare variants as more patient histologies and sequences become available in the future.

### 3.2.3 SPOP mutation state prediction is accurate and the prediction uncertainty is bounded within a confidence interval

Our pipeline's accuracy in SPOP mutation state prediction is statistically significant: (i) within held-out test datasets in TCGA, (ii) within held-out test datasets in MSK-IMPACT, (iii) from TCGA against the independent MSK-IMPACT cohort (Fig 3.5), and (iv) vice versa. The confidence interval [CI] calculated using seven ResNet ensembles allows every prediction to be evaluated statistically: is there significant evidence for SPOP mutation in the patient, is there significant evidence for SPOP non-mutation in the patient, or is the patient's SPOP mutation state inconclusive. Clinicians may choose to follow the mutation prediction only if the histological imagery provides acceptably low uncertainty.

### 3.2.4 SPOP mutation state prediction is automated and does not rely on human interpretation

Unlike Gleason score, which relies on a clinician's interpretation of the histology, our pipeline is automated (Fig ) – though detecting overstain, blur, and mostly background in the slide are not yet automated (Alg 3.1). The pipeline's input is the whole digital slide and the output is the SPOP mutation prediction bounded within 95% and 99% CIs. Moreover, our pipeline does not require a human to identify a representative region in the slide, as is done to create tissue microarrays [TMAs] from slides.

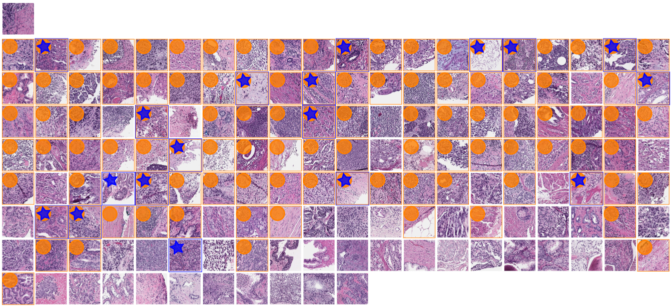


Figure 3.6: **CBIR performance** SPOP mutant query image at top left for CBIR using TCGA-trained metaensemble on MSK-IMPACT dataset, showing most similar images in top row, with most similar image leftmost. Less similar images in lower rows, ordered left to right by similarity. One image patch per patient shown. Though the full octagon of patches (see Fig 3.4B, for query) is considered for the query and each retrieved patient, only the center patch is shown here, with a blue surrounding box and star indicating SPOP mutation, while an orange box and circle indicates this retrieved patient's predicted SPOP mutation 95% CI overlaps with that of the query patient. A 95% CI may in practice usefully limit the number of retrieved results. The most similar patient (top row, leftmost), second most, third, fourth, and fifth have dissimilarity scores of 0.073936, 0.092472, 0.107029, 0.108397, and 0.109131, respectively. The least similar patient (bottom right) has dissimilarity score 0.452150. Mean dissimilarity is 0.219295 with standard deviation 0.075968. Similarity is 1 - dissimilarity. The second most similar patient (top row, second from left) is a mutant, like the query. There are 5 mutants in the top row, 3 in the second, etc, with zero in the bottom row. Patients with SPOP mutation are more similar than patients without SPOP mutation for this query (p=0.02032, one-tailed Mann-Whitney U test).

### 3.2.5 SPOP mutation state prediction finds patients that appear to share the same genetic driver of disease

Each ensemble (Fig   ) may be used as a feature to predict patient similarity, where similar patients on average share the same ensemble-predicted SPOP mutation probability (Fig 3.6). A patient's histology may be studied in the context of similar patients' for diagnostic, prognostic, and theragnostic considerations.

### 3.2.6 Molecular pathology, such as characterizing histology in terms of SPOP mutation state, leads directly to precision medicine

For instance, non-mutant SPOP ubiquitinylates androgen receptor [AR], to mark AR for degradation, but mutant SPOP does not. Antiandrogen drugs, such as flutamide and enzalutamide, promote degradation of AR to treat the cancer, though mutant AR confers resistance[65,66].

### 3.2.7 SPOP mutation state prediction provides information regarding other molecular states

SPOP mutation is mutually exclusive with TMPRSS2-ERG gene fusion[56], so our SPOP mutation predictor provides indirect information regarding the TMPRSS2-ERG state and potentially others.

## 3.3 Discussion

We summarize a whole slide with a maximally abnormal subregion within the dominant tumor, such that the vast majority of the slide is not used for deep learning. Tissue microarrays take a similar though manual approach, using a small circle of tissue to represent a patient's cancer. In a sense, our patch extraction, saliency prediction, and TMARKER-based cell counting pipeline stages together model how a representative patch may be selected to identify the dominant cancer subtype in the slide overall, ignoring other regions that may not have the same genetic drivers. This subregion has many desirable properties: **(i)** it is salient at low magnification, i.e. diagnostically relevant to a pathologist at the microscope[67], **(ii)** it has the maximum number of malignant cells at low magnification, which is a heuristic to locate the dominant tumor, presumably enriched for cells with driver mutations such as SPOP due to conferred growth advantages and reflected in the bulk genetic sequencing we use as mutation ground truth, **(iii)** it has the maximum number of abnormal cells at high magnification, which is a heuristic to locate a subregion with most nuanced cellular appearance, presumably concentrated with cellular-level visual features that can discriminate between cancers driven by SPOP versus other mutations, and **(iv)** at 800x800 pixels, it is large enough at high magnification to capture histology, e.g. gland structure.

A holistic approach that considers for deep learning patches spatially distributed widely throughout the slide, rather than only the dominant tumor, could improve performance and provide further insight. Gleason grading involves such a holistic approach, identifying first and second most prevalent cancer morphologies in a whole slide. The complexities of multiple and varying counts of representatives per slide are future work.

We use multiple classifier ensembles to predict the SPOP mutation state. Though each residual network[62] classifier tends to predict SPOP mutation probability in a bimodal distribution, i.e. being either close to 0 or close to 1, averaging these classifiers within an ensemble provides a uniform distribution representing the SPOP mutation probability (Fig S3.8).

Deep learning typically requires large sets of training data, yet we have only 20 patients with SPOP mutation, nearly an order of magnitude fewer than the 157 patients without somatic SPOP mutation in the TCGA cohort. Deep learning in small data is a challenging setting. We confront this challenge by training many residual networks [ResNets] on small draws of the data (Alg 3.2), in equal proportions of SPOP mutants and SPOP non-mutants, then combining the ResNets as weak learners, to produce a strong learner ensemble[68,69], similar in principle to a random forest ensemble of decision trees[70]. Empirically, the great depth of ResNets appears to be important because CaffeNet[49] – a shallower 8-layer neural network based on AlexNet[50] – so rarely achieved validation accuracy of 0.6 or more predicting SPOP mutation state than an ensemble could not be formed.

Pretraining the deep networks is essential in small data regimes such as ours, and we use the ResNet-50 model pretrained on ImageNet[62]. We used both the published ResNet-50 and our customized ResNet-50 that included an additional 50% dropout[48] layer and a 1024-neuron fully-connected layer. In practice, for difficult training set draws, often at least one of these architectures converged to validation accuracy of 0.6 or more. For data augmentation, the 800x800px images at high magnification were trimmed to the centermost 512x512px in 6 degree rotations, scaled to 256x256px, flipped and unflipped, then randomly cropped to 224x224 within Caffe[49] for training. Learning rates varied from 0.001 to 0.007, and momentums from 0.3 to 0.99. Test error tended to be worse with momentum less than 0.8, though low momentum allowed convergence for difficult training sets. Learning rates of 0.001, 0.0015, or 0.002 with momentum of 0.9 was typical. We used nVidia Titan-X GPUs with 12GB RAM for training. At 12GB, one GPU could train either ResNet architecture using a minibatch size of 32 images.

## 3.4 Materials and methods

Please see Section S3.5 for discussion of (i) digital slide acquisition and quality control, (ii) region of interest identification, (iii) data augmentation, (iv) deep learning, and (v) cross-validation.

### 3.4.1 Ensemble aggregation

To estimate generalization error, the 11 classifiers in a trial were selected by highest validation accuracy to form an ensemble, and a metaensemble from 7 independent trial ensembles. A 95% basic bootstrapped CI[3] indicated generalization accuracy from these 7 trials was 0.58-0.86 on TCGA (Fig 3.5A) and 0.61-0.83 on MSK-IMPACT (Fig 3.5B), – both significantly better than 0.5 chance. On TCGA, metaensemble AUROC was 0.74 (p=0.00024), accuracy was 0.70, and Fisher's Exact Test p=0.00070 (Fig 3.5C1). On MSK-IMPACT, metaensemble AU-

---

[3] We used the R `boot` library for basic bootstrap CIs. Canty and Ripley 2016: `https://cran.r-project.org/web/packages/boot/index.html`

ROC was 0.75 (p=0.00017), accuracy was 0.73, and Fisher's Exact Test p=0.00010 (Fig 3.5D1). These two single-dataset estimates indicate the method performs significantly better than chance on unseen data. We formed tuned ensembles (Fig S3.10), where 11 classifiers in a trial were selected by highest ensemble *test* accuracy, which no longer estimates generalization error but promotes correct classification on average across ensembled classifiers for each test example (Fig 3.5A,B white versus gray at right).

### 3.4.2 Independent evaluation

We tested the TCGA-trained metaensemble of 7 tuned 11-classifier ensembles on the MSK-IMPACT cohort, and vice-versa. Each patient had one slide. For each slide, a central region of interest and surrounding octagon of patches were found, for nine patches per slide (Fig 3.4B,C). We hypothesized that SPOP mutation leads to 1-2 nine-patch-localized lesions, meaning adjacent patches have similar classifier-predicted SPOP mutation probability. Therefore, we first calculated the SPOP mutation prediction weighted mean over all classifiers in an ensemble, with classifiers having greater nine-patch prediction variance being more weighted in the mean. Second, of the nine patches we formed all possible groups of three adjacent patches and assigned to a group the minimum SPOP mutation prediction weighted mean of any patch in the grouped three, then took the second-greatest group as a classifier's overall patient cancer SPOP mutation prediction. Within an ensemble, we weighted a classifier's overall prediction by the classifier's nine-patch mutation prediction variance, because classifiers with high variance (e.g. mostly mutant, but 1-3 of nine non-mutant) reflect the lesion hypothesis better than classifiers that predict all non-mutant

(0) or all mutant (1). Thus on MSK-IMPACT, the TCGA metaensemble assigned stochastically greater scores to mutants than non-mutants (AUROC 0.86) and accurately distinguished mutants from non-mutants (Fisher's Exact p=0.00379) (Fig 3.5C2). Moreover on TCGA, the MSK-IMPACT metaensemble assigned stochastically greater scores to mutants than non-mutants (AUROC 0.64) and accurately distinguished mutants from non-mutants (Fisher's Exact p=0.03056) (Fig 3.5D2).

### 3.4.3 CBIR evaluation

The metaensemble consists of seven ensembles. Each ensemble predicts SPOP mutation probability as a uniformly distributed random variable (Fig S3.8). Each ensemble is tuned to a different test set (Sec S3.5.6 and Table S3.1), so we treat each ensemble's prediction as a feature, for seven 32-bit SPOP CBIR features total. For CBIR, a patient's dissimilarity to the query patient is the mean of absolute differences in these seven features, e.g. a dissimilarity of 0.2 means on average an ensemble predicts the patient's SPOP mutation probability is 0.2 different than the query. A similarity score is 1 minus dissimilarity. We evaluated the TCGA-trained metaensemble on each of the 19 SPOP mutants in MSK-IMPACT (Fig 3.1), for each mutant calculating a CBIR p-value, with a low p-value indicating it is not due to chance alone that dissimilarities of mutants were lower than non-mutant dissimilarities (Fig 3.6). To conduct a metaanalysis of these 19 dependent p-values, we may use neither (a) Fisher's Method[71] because our p-values are not independent, nor (b) Empirical Brown's Method[72] because 100 p-values are required for convergence, so we used Kost's

Method[4] [73]. Kost's Method found an overall p=0.024076 (14.6 degrees of freedom [d.f.]), so it is not due to chance alone that our CBIR returns SPOP mutant patients with lower dissimilarity scores than non-mutants for each SPOP mutant patient query. Similarly, our CBIR tool returned non-mutant patients with lower dissimilarity than SPOP mutants for each of the 133 MSK-IMPACT non-mutant patients (p=0.016960 Kost's Method, 22.2 d.f.).

## Acknowledgements

AJS was supported by NIH/NCI grant F31CA214029 and the Tri-Institutional Training Program in Computational Biology and Medicine (via NIH training grant T32GM083937). TJF was supported by an MSK Functional Genomics Initiative grant. This research was funded in part through the NIH/NCI Cancer Center Support Grant P30CA008748. We gratefully acknowledge NVIDIA corporation for GPU support under the GPU Research Center award to TJF. TJF is a founder, equity owner, and Chief Scientific Officer of Paige.AI.

## S3.5 Supplementary materials and methods

### S3.5.1 Digital slide acquisition and quality control

TCGA slides and SPOP mutation state was downloaded from cBioPortal [74]. The MSK-IMPACT cohort was available internally. We discarded slide regions that were not diagnostically salient[67]. We discarded slide regions that were (i) overstained eosin blue, (ii) $>$ 50% erythrocytes, (iii) $>$ 50% empty, or (iv) $>$ 50% blurred, by inspection (Alg 3.1). Patients having no regions remaining were discarded. A region is a 800x800 pixel [px] patch at 8 microns per pixel [$\mu pp$], ˜100x magnification.

---

[4] We used Kost's Method from R package EmpiricalBrownsMethod https://www.bioconductor.org/packages/devel/bioc/vignettes/EmpiricalBrownsMethod/inst/doc/ebmVignette.html

### S3.5.2 Region of interest identification

For each region, we used TMARKER[75] to identify cell nuclei and classify them as benign, malignant, or unknown – using a corresponding 12750x12750px patch at level 0 of the slide SVS file (0.25 $\mu pp$) for the 800x800px region passing quality control (Fig ). The 12750x12750px patch having the most malignant cells in this slide was deemed the "dominant tumor" after searching in a 10px grid for an 800x800px region of interest patch (at 4 $\mu pp$) having maximal abnormal cells. A malignant cell is 1 abnormal cell while an unknown cell is 0.5 abnormal cell and 0.5 normal cell. We expected a mutant driver gene, such as SPOP, to confer a growth advantage evident in the region of interest and show histological features, such as ducts. By maximizing abnormal cell counts, we intended to find as many discriminative image textures based on SPOP-mutation-driven cancer cells as possible.

### S3.5.3 Data augmentation

In an 800x800px region of interest patch, the centermost 512x512px patch was selected, rotated in six degree increments for rotational invariance, and scaled to 256x256px. Eight 800x800px patches were also selected in a surrounding octagon formation around the region of interest, offset 115 and 200px away from the central region of interest, so two adjacent octagon patches were 230px away from each other and 230px away from the central patch (Fig 3.4B,C). These patches were rotated and scaled. This integrates a circular area to summarize the slide, akin to circular tissue cuts for TMA spots.

### S3.5.4 Deep learning

We used Caffe[49] for deep learning a binary classification model given the 256x256px patches labeled by SPOP mutation state. We adapted the ResNet-50[62] model pre-trained on ImageNet[51] data by re-initializing the last layer's weights (Fig 3.4A), then trained on pathology images. Classifier predictions followed a sharply bimodal distribution of 0 or 1, but ensemble predictions followed a uniform distribution suitable for CIs (Fig S3.8).

### S3.5.5 Cross-validation

For a trial, a test set of 5 positive and 5 negative [5+/5-] examples was used, where a positive example is a patient with SPOP mutation (Table S3.1). If a patient had multiple slides that passed quality control, at most one was used in the trial, taken at random. With the test set held out, we ran Monte Carlo cross-validation 13 times, each time (i) drawing without replacement a training set of 10+/10- examples and validation set of 5+/5- examples, and (ii) training both a ResNet-50 and Drop50 classifier (Fig 3.4A). Our convergence criterion was classifier validation accuracy >= 0.6 with 10000 – 100000 training iterations. An ensemble was 11 classifiers, one per Monte Carlo cross-validation, allowing 2 of 13 Monte Carlo cross-validation runs to converge poorly or not at all (Alg 3.2). Empirically, this class-balanced stratified-sampling ensemble approach reduces classifier bias towards the majority class and broadly samples both classes.

```
input : prostate adenocarcinoma patients: with whole slide images and
        genomic sequencing
        saliency(...): a whole slide image region saliency predictor
        cancer(...), abnormal(...): cell nuclei classifiers
output: data set: of maximally abnormal patches and SPOP state
for patient ⟵ prostate adenocarcinoma patients do
    spop ⟵ SPOP normal/mutated state as 0/1;
    slide ⟵ whole slide image of cancer biopsy;
    patches ⟵ 75% overlap 800x800px images ⟵ slide;
    salient patches ⟵ predict saliency(patches);
    dom tumor patch ⟵ max cancer(salient patches);
    abn patch ⟵ max abnormal(dom tumor patch);
    if abn patch < 50% blurred then
        if abn patch < 50% background then
            if abn patch < 50% erythrocytes then
                if abn patch not eosin overstained then
                    append (abn patch, spop) to data set;
                end
            end
        end
    end
end
return data set;
```

**Algorithm 3.1: Dominant tumor cell counting as slide preprocessing.** Each patient's SPOP mutation state is paired with the whole slide image patch having the maximum number of abnormal cells, where abnormal is either cancer or unknown cell type, rather than healthy cell type. This patch is taken from the dominant tumor patch. The dominant tumor patch is both salient and has the maximum number of cancer cells. The dominant tumor patch is at low magnification, while the abnormal patch is at high magnification within the dominant.

```
slide ← whole slide image of cancer biopsy;
patches ← 75% overlap 800x800px images ∈ slide;
salient_patches ← predict_saliency(patches);
dom_tumor_patch ← max_cancer(salient_patches);
abn_patch ← max_abnormal(dom_tumor_patch);
for ensemble ∈ ensembles do
    $I_0, I_1, \ldots, I_8$ ← abn_patch and surrounding octagon;
    spop_prediction ← $P_{indep}(X|I_0, I_1, \ldots, I_8, C_1, C_2, \ldots, C_{11})$;
    append spop_prediction to spop_predictions;
end
return confidence_interval(spop_predictions);
```

Algorithm 3.3: Metaensemble SPOP mutation prediction. Use each of 7 ensembles to predict SPOP mutation state, then compute a bootstrapped CI of SPOP mutation state given these predictions. Each prediction is a probability of SPOP mutation in the patient. If this probability is sufficiently high to surpass a defined threshold, the patient's cancer is predicted to have an SPOP mutation. Likewise, if this probability is sufficiently low to fall below a defined threshold, the patient's cancer is predicted to not have an SPOP mutation relative patient germline. Probabilities between these upper and lower thresholds are inconclusive predictions. Threshold calibration depends both on the dataset and the cost of false positives/negatives. For instance, if the CI lower bound is $> 0.5$, there is significant confidence that the patient has SPOP mutation. If the CI upper bound is $< 0.5$, there is significant confidence that the patient does not have SPOP mutation. Otherwise, the patient's SPOP mutation state cannot be confidently predicted. When training on TCGA and testing on MSK-IMPACT with a 99% CI (Fig 3.5C3), the metaensemble did not commit any false negatives at a 0.5 cutoff, so we calibrated the cutoff to 0.495 to make one more true positive while still not committing false negatives. Thus SPOP mutation was predicted when the lower CI bound was above 0.495, and SPOP was predicted to not be mutated when the upper CI bound was below 0.495, with the remaining cases being inconclusive (Fig 3.5C2,C3). When training on MSK-IMPACT and testing on TCGA, false positives were an issue, presumably because TCGA frozen sections are poorer image quality than MSK-IMPACT fixed sections (Fig 3.1), so we changed the lower threshold to the mean, converting many false positives to true negatives. Thus SPOP mutation was predicted when the lower CI bound was above 0.495, and SPOP was predicted to not be mutated when the metaensemble mean was below 0.495 (Fig 3.5D2,D3), with the remaining cases being inconclusive. In this way, the CI bounds or mean served as thresholds in a calibrated dataset-dependent manner, and we recommend such calibration when testing on a new dataset. Section S3.5.7 defines $P_{indep}(X|I_0, I_1, \ldots, I_8, C_1, C_2, \ldots, C_{11})$, which calculates an overall SPOP mutation prediction for the patient given the central patch, surrounding octagon of patches (Fig 3.4B,C), and 11 classifiers in an ensemble.

### S3.5.6 TCGA dataset learning

See Fig 3.5C1 for TCGA testing of the TCGA-trained metaensemble, with Fig S3.10 showing TCGA-trained tuned metaensemble performance. Training was as follows:

1. $P_{learn}(X = Mutant) = P_{learn}(X = Normal) = 0.5$. Each classifier is trained on twenty patients, with ten having mutant SPOP with respect to the patient germline. With this balanced dataset for training, the probability that a training example is a mutant is the same as the probability that a training example is normal, 0.5. This differs from the proportion of patients having a somatic SPOP mutation, ~10% in the TCGA dataset. Classifiers that learn on unbalanced datasets may simply learn to predict the predominant class.
2. Learning $P_{learn}(X|I, C_1, C_2, ..., C_{11})$ for each patient representative histology image I and eleven classifiers C. This learns the probability that a classifier will predict the patient is mutated, given the representative histology image of the patient. An ensemble of an infinite number of classifiers would give the expected value of this probability. We use eleven classifiers. Each classifier tends to emit either a zero or one, predicting mutant or normal, i.e. $P_{learn}(X|I, C)$ tends to be 0 or 1. We take the ensemble average to be the probability $P_{learn}$, which tends to be uniformly distributed rather than bimodal (Fig S3.8): $P_{learn} = P_{learn}(X|I, C_1, C_2, ..., C_{11}) = (1/11) * \sum_{i=1}^{11} P_{learn}(X|I, C_i)$.
3. Training and validation sets consist of a central image patch and eight surrounding image patches in an octagon formation, where the distance

between the central patch and a surrounding patch is the same as the distance between any two adjacent surrounding patches. All patches had the same label, either 1 for patients with somatic SPOP mutation or 0 for patients without somatic SPOP mutation. This approximately circular region of nine patches per patient provides more information for training, more closely resembles the region available from a circle of tissue in a TMA, and does not sample far away tissues that may have different molecular drivers of disease.

4. The test set consists of a single patch per patient, without a surrounding octagon of eight patches. Test accuracy and squared loss were optimized from the converged trained models having at least 0.6 validation accuracy, i.e. the "sel" ensembles in Fig S3.10. This optimization selected at most one model from 11 of 13 Monte Carlo cross validation runs, where the selection of 11 models for the ensemble had the highest accuracy and lowest squared loss on the test set. This (a) ensures on unseen data the ensemble has both the discriminative power and diversity to correctly predict $P_{learn}$, which is a function of a single image, and (b) optimizes each ensemble according to a different test set, with the aim of decorrelating ensembles for a more general metaensemble, much like how decision trees should be decorrelated for a more general random forest[70].

### S3.5.7 MSK-IMPACT dataset testing

See Fig 3.5C2,C3 for MSK-IMPACT testing of the TCGA-trained metaensemble. Testing was as follows, and the analogous procedure holds for TCGA testing of the MSK-IMPACT-trained metaensemble (Fig 3.5D2,D3):

1. $9 * P_{indep}(X = Mutant) \approx P_{indep}(X = Normal)$. The ensembles learned on the TCGA dataset are tested on an independent MSK-IMPACT dataset 152 patients, 19 having somatic SPOP mutation, so the probability that one of these MSK-IMPACTed patients is mutated is only ˜10%. This is in sharp contrast to the TCGA training set, where the probabilities of mutant and non-mutant patients are equal, due to the stratified sampling for training. Because $P_{learn}(X = mutant) = 0.5$ is greater than $P_{indep}(X = mutant) \approx 0.1$, additional scaling or stringency is required to avoid false positives from the learned classifier applied to the independent dataset.

2. Evaluating $P_{indep}(X|I_0, I_1, ...I_8, C_1, C_2, ..., C_{11})$ for each patient. Each of the eleven classifiers are evaluated on the patient's central representative histology image and the octagon of eight surrounding images (Fig 3.4B,C), to predict SPOP mutation. Because each classifier has been trained on only 10 mutants and 10 non-mutants, many classifiers will emit a biased prediction of 0 for all nine images or 1 for all nine images, lacking power to discriminate among the nine images. Empirically however, 1-3 classifiers tend to have greater variance in their predictions for each of the nine images of the patient. The classifiers with greater variance tend to correctly predict SPOP mutation state, presumably because the features learned from similar training data can distinguish image regions showing mutant image textures from regions showing non-mutant image textures. Such mutant and non-mutant regions should spatially cluster, i.e. a classifier should make the same SPOP mutation state prediction for adjacent images in a patient, otherwise the classifier is predicting noise. Moreover, classifiers with greater variance should agree which patches have high evidence of mutation and which patches have low evidence of mutation.

Therefore, we take the classifier mean prediction, for all 11 classifiers in the ensemble, weighting more highly in this mean the predictions from classifiers with higher interpatch prediction variance. We define the weighted ensemble average as $P_{indep}(X|I, C_1, C_2, ..., C_{11}) = (\sum_{i=1}^{11} Var_i * P_{learn}(X|I, C_i)) / \sum_{i=1}^{11} Var_i$, where we define $Var_i = (\sum_{j=0}^{8}(P_{learn}(X|I_j, C_i) - \mu_i)^2) + \frac{(1-\mu_i)^2+(0-\mu_i)^2}{8}$ and $\mu_i = (1/11)\sum_{k=1}^{11} P_{learn}(X|I_k, C_i)$, so that even classifiers that predict 0 for all images or 1 for all images still have some small weight. If, for instance, the mean predictions ($P_{indep}(X|I, C_1, C_2, ..., C_{11})$) suggest SPOP is mutated in each of 4 adjacent images, there is significant evidence for SPOP mutation at $\alpha = 0.05$: of 9 images, 16 possibilities exist of 4 adjacent images being SPOP mutants and the other images being non-mutants, and $2^9 = 512$ mutant/non-mutant configurations possible, so probability of 4 adjacent images is $16/512 = 0.03125 < 0.05 = \alpha$. Thus for added stringency, one may take as the overall corrected $P_{indep}$ the maximum of all adjacent 4-image cluster mean prediction minima: $P_{indep4}(X|I_0, I_1, ...I_8, C_1, C_2, ..., C_{11}) = \max(\min_{i \in a,b,c,d}(P_{indep}(X|I_i, C_1, C_2, ..., C_{11})))$. However, due to the noisy nature of the mean predictions from classifiers trained on little data, a mutant 4-cluster may not occur, but additional adjacent images indicating mutation may offer supporting evidence. From the observation that a 4-cluster is two overlapping 3-clusters, we slightly loosen the correction stringency to $P_{indep3pen}(X|I_0, I_1, ...I_8, C_1, C_2, ..., C_{11}) = pen(\min(P_{indep}(X|I_a, C_1, C_2, ..., C_{11}), P_{indep}(X|I_b, C_1, C_2, ..., C_{11}), P_{indep}(X|I_c, C_1, C_2, ..., C_{11})))$, $\forall \{I_a, I_b, I_c\} \in \{I_0, I_1, ..., I_8\}$ $s.t$ $I_a \, adjacent \, I_b$, $I_b \, adjacent \, I_c$, where $pen(...)$ returns the second-greatest value, or penultimate. In this way, $P_{indep3pen}$ is identical to $P_{indep4}$ for a 4-cluster, but additionally allows partially- or non-overlapping 3-clusters to suggest SPOP mutation. $P_{indep3pen}$ is strictly greater than or equal to $P_{indep4}$.

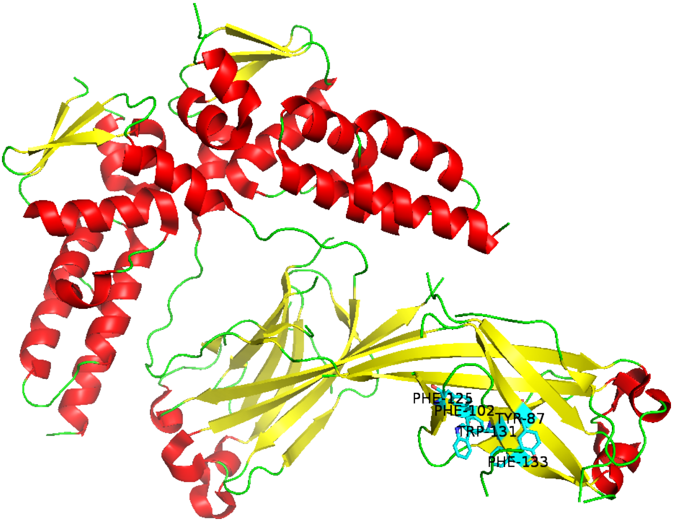


Figure S3.7: **SPOP tertiary structure with mutated residues** labeled, common to one site. Though most genetic aberrations affect the substrate-binding MATH domain in SPOP, our method's performance appears robust to two of 20 patients having genetic changes outside this pocket. Mutation chr17:47677762 R368H in patient TCGA-EJ-7782-01 not shown and not in site. Deletion chr17:47699392 in patient TCGA-VP-A878 not shown and not in site. PDB structure 3HQI.

### S3.5.8 Implementation details

We studied a TCGA cohort of 499 prostate adenocarcinoma patients, in particular the 177 patients that had both acceptable pathology images and SPOP mutation state available (Fig 3.1, Alg 3.1). SPOP mutation state was downloaded from cBioPortal[74]. After learning on TCGA data, we tested on an MSK-IMPACT dataset of 19 SPOP mutant patients and 119 non-mutant patients.

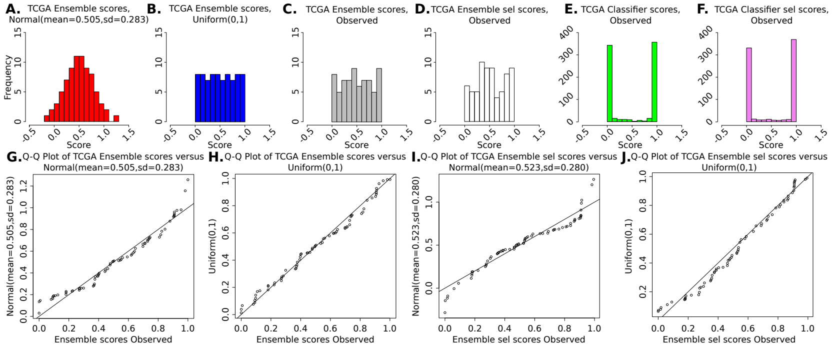
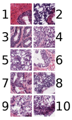
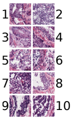


Figure S3.8: **TCGA classifier and ensemble mutation prediction scores**, showing ensemble scores are uniform random variables representing the mutation probability when the mutant and non-mutant classes are equally sampled, despite the sharply bimodal score distribution of individual classifiers (Table S3.1). *Compared to the bimodal distribution of single classifiers, this uniform random distribution enables valid confidence interval calculation over ensemble mutation prediction scores* and additionally does not suffer confounds such as p-value inflation, e.g. a disproportionate number of ensemble mutation prediction scores being close to zero. **Panel A:** Normal distribution of scores with mean and stdev of TCGA-trained ensemble prediction scores, showing the distribution domain extends beyond the valid [0,1] domain of mutation prediction scores. **Panel B:** Uniform distribution of scores, with all scores in the valid [0,1] domain and none outside. **Panel C:** TCGA-trained ensemble scores, following Uniform distribution (KS Test p=0.9792). Corresponds to Fig 3.5A, second from far right. **Panel D:** TCGA-trained tuned ensemble scores, following a Uniform distribution (KS Test p=0.4687). Corresponds to Fig 3.5A, far right. The excess of predictions around 0.5 is an artifact of the tuning process, which tends to select classifiers for the ensemble such that the ensemble prediction is at least 0.5 for mutants and less than 0.5 for non-mutants if the ensemble prediction was incorrect before tuning. See also Fig S3.10 for tuned ensemble performance. **Panel E:** TCGA-trained classifier scores, following sharply bimodal distribution of 0 (non-mutant) or 1 (mutant) predictions. **Panel F:** TCGA-trained classifiers selected for the tuned ensemble are also bimodal. **Panel G:** Q-Q plot showing outliers, indicating TCGA-trained ensemble scores do closely follow a Normal distribution. **Panel H:** Q-Q plot showing close linear relationship, indicating TCGA-trained ensemble scores follow a Uniform distribution. **Panel I:** Q-Q plot showing outliers, indicating TCGA-trained tuned ensemble scores do not closely follow a Normal distribution. **Panel J:** Q-Q plot showing close linear relationship, indicating TCGA-trained tuned ensemble scores follow a Uniform distribution.

Table S3.1: **TCGA test set images and accuracies**, with prediction errors from single residual networks highlighted in red and prediction errors from residual network ensembles highlighted in magenta. New 20-patient training and 10-patient validation sets are drawn for each Monte Carlo cross validation run against the trial's 10-patient test set. The far right column is used to calculate the mean classifier prediction, which is used as the trial's ensemble prediction. The Receiver Operating Characteristic of these ensembles in shown in Fig S3.9. The seven ensemble predictions are used to calculate a confidence interval of generalization accuracy, shown in Fig 3.5A second from far right in gray and red. Individual classifier predictions follow a sharply bimodal distribution of 0 (non-mutant) or 1 (mutant), while the ensemble predictions follow a uniform random distribution (Fig S3.8). Table S3.1 is continued in Table S3.2, showing the final two enembles in the seven ensemble metaensemble trained on TCGA data.

| Trial0 | Test # : SPOP | Monte1 | Monte2 | Monte3 | Monte4 | Monte5 | Monte6 | Monte7 | Monte8 | Monte9 | Monte10 | Monte11 | Mean |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 : 0 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.9988 | 0.0000 | 0.0000 | 0.9987 | 0.1816 |
| | 2 : 1 | 1.0000 | 1.0000 | 1.0000 | 0.8637 | 0.0613 | 1.0000 | 0.1741 | 0.1386 | 1.0000 | 1.0000 | 0.9647 | 0.7457 |
|  | 3 : 0 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0001 | 0.0000 | 1.0000 | 0.0000 | 0.9892 | 0.0000 | 0.1808 |
| | 4 : 1 | 0.3284 | 0.0000 | 0.0000 | 0.7208 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 0.9691 | 1.0000 | 0.7289 |
| | 5 : 0 | 1.0000 | 0.0006 | 0.0001 | 0.0020 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 0.9845 | 0.0001 | 0.0114 | 0.5453 |
| | 6 : 1 | 0.0124 | 1.0000 | 0.9970 | 0.2412 | 0.9999 | 0.9982 | 0.0000 | 0.0001 | 0.0017 | 0.0000 | 0.9990 | 0.4772 |
| | 7 : 0 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0750 | 0.0000 | 0.9931 | 0.0001 | 0.0000 | 0.0000 | 0.0971 |
| | 8 : 1 | 0.1069 | 1.0000 | 0.9992 | 0.8635 | 0.0054 | 1.0000 | 0.9991 | 1.0000 | 0.6994 | 0.9997 | 0.0036 | 0.6979 |
| | 9 : 0 | 0.9990 | 0.0156 | 0.9989 | 0.0005 | 0.9978 | 0.9844 | 1.0000 | 0.1079 | 0.4288 | 0.0000 | 0.0000 | 0.5030 |
| | 10 : 1 | 0.0013 | 0.2339 | 0.9962 | 0.0001 | 0.9978 | 0.0000 | 1.0000 | 1.0000 | 0.8206 | 0.9975 | 0.0197 | 0.5516 |

| Trial1 | Test # : SPOP | Monte1 | Monte2 | Monte3 | Monte4 | Monte5 | Monte6 | Monte7 | Monte8 | Monte9 | Monte10 | Monte11 | Mean |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 : 0 | 0.1674 | 0.0007 | 0.0041 | 0.0005 | 0.6145 | 0.0001 | 0.0002 | 0.0016 | 0.3730 | 0.9999 | 0.0000 | 0.1813 |
| | 2 : 1 | 0.9580 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 0.9973 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 0.9542 |
|  | 3 : 0 | 0.0839 | 0.0000 | 0.0001 | 0.0000 | 0.0000 | 0.0048 | 0.0000 | 0.0022 | 0.0000 | 1.0000 | 0.0000 | 0.0916 |
| | 4 : 1 | 0.0845 | 0.0000 | 0.0000 | 0.0000 | 0.0008 | 0.0000 | 0.0000 | 0.0114 | 0.0000 | 1.0000 | 0.0000 | 0.0920 |
| | 5 : 0 | 0.6265 | 0.9999 | 1.0000 | 0.9998 | 0.9747 | 0.9933 | 0.0001 | 1.0000 | 0.0000 | 0.3829 | 0.1641 | 0.5925 |
| | 6 : 1 | 0.4271 | 0.1278 | 1.0000 | 0.0008 | 0.0010 | 0.0000 | 0.0000 | 1.0000 | 0.8001 | 1.0000 | 0.0003 | 0.3750 |
| | 7 : 0 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| | 8 : 1 | 0.4438 | 0.0000 | 1.0000 | 0.0002 | 0.4147 | 0.0000 | 0.9095 | 0.0011 | 1.0000 | 1.0000 | 0.0000 | 0.3932 |
| | 9 : 0 | 0.4236 | 0.0000 | 0.3012 | 0.0004 | 1.0000 | 0.9998 | 0.0000 | 0.0001 | 0.8132 | 1.0000 | 0.0000 | 0.4621 |
| | 10 : 1 | 0.8278 | 0.9999 | 1.0000 | 1.0000 | 1.0000 | 0.9312 | 1.0000 | 1.0000 | 0.9999 | 0.9729 | 1.0000 | 0.9030 |

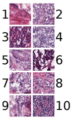
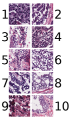
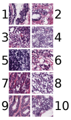
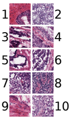
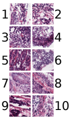
| Trial2 | Test # : SPOP | Monte1 | Monte2 | Monte3 | Monte4 | Monte5 | Monte6 | Monte7 | Monte8 | Monte9 | Monte10 | Monte11 | Mean |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 : 0 | 0.0744 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0002 | 0.0001 | 0.0000 | 0.0000 | 0.0077 | 0.0075 |
| | 2 : 1 | 0.9993 | 0.0022 | 0.9999 | 0.1242 | 1.0000 | 0.0000 | 0.0412 | 1.0000 | 0.9674 | 0.9995 | 0.9880 | 0.6474 |
|  | 3 : 0 | 0.2742 | 0.0000 | 0.8778 | 0.0000 | 1.0000 | 0.0044 | 1.0000 | 0.1288 | 0.0040 | 0.0000 | 0.0000 | 0.2990 |
| | 4 : 1 | 1.0000 | 0.9989 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 0.0008 | 1.0000 | 0.9091 |
| | 5 : 0 | 0.0000 | 0.0057 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0019 | 0.0002 | 0.0000 | 0.0000 | 0.0007 |
| | 6 : 1 | 0.9904 | 0.0003 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 0.9994 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 0.9082 |
| | 7 : 0 | 0.9971 | 0.0002 | 0.0001 | 0.9961 | 1.0000 | 0.0000 | 0.9309 | 0.0000 | 0.0000 | 0.0019 | 0.9975 | 0.4476 |
| | 8 : 1 | 0.9968 | 0.0001 | 0.2478 | 0.0057 | 1.0000 | 0.0000 | 1.0000 | 0.0037 | 0.0000 | 0.9858 | 0.2014 | 0.4038 |
| | 9 : 0 | 0.9628 | 0.4937 | 0.3145 | 0.0126 | 0.9987 | 0.6667 | 0.0002 | 0.0721 | 0.0000 | 0.0000 | 0.6324 | 0.3776 |
| | 10 : 1 | 1.0000 | 0.0000 | 1.0000 | 0.7263 | 1.0000 | 1.0000 | 0.9654 | 1.0000 | 0.0043 | 1.0000 | 0.4465 | 0.7402 |

Table S3.2: Continuation of Table S3.1, for trials 3, 4, 5 and 6, the final four ensembles in the TCGA-trained metaensemble.

| Trial3 | Test # : SPOP | Monte1 | Monte2 | Monte3 | Monte4 | Monte5 | Monte6 | Monte7 | Monte8 | Monte9 | Monte10 | Monte11 | Mean |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
|  | 1 : 0 | 0.9820 | 0.0000 | 0.0000 | 0.0000 | 0.9999 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 1.0000 | 0.2711 |
| | 2 : 1 | 0.0703 | 0.3796 | 0.0034 | 1.0000 | 0.0007 | 1.0000 | 1.0000 | 0.0000 | 0.9955 | 0.9872 | 0.0364 | 0.4976 |
| | 3 : 0 | 0.9994 | 1.0000 | 0.1613 | 0.0000 | 0.9959 | 0.2808 | 1.0000 | 1.0000 | 0.0010 | 0.0003 | 0.0000 | 0.4944 |
| | 4 : 1 | 1.0000 | 0.0002 | 1.0000 | 0.3774 | 0.9294 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 0.8461 |
| | 5 : 0 | 0.0000 | 0.0419 | 0.0016 | 0.0009 | 0.0000 | 1.0000 | 1.0000 | 1.0000 | 0.0003 | 0.8270 | 1.0000 | 0.4429 |
| | 6 : 1 | 1.0000 | 1.0000 | 0.0104 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 0.0989 | 1.0000 | 1.0000 | 0.0011 | 0.7373 |
| | 7 : 0 | 0.0688 | 0.9645 | 0.0434 | 1.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 1.0000 | 0.9813 | 0.0009 | 0.3690 |
| | 8 : 1 | 0.9998 | 1.0000 | 0.8986 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 0.9981 | 1.0000 | 0.0004 | 0.0000 | 0.8088 |
| | 9 : 0 | 0.0002 | 0.0147 | 1.0000 | 0.0000 | 0.5648 | 0.0002 | 0.0000 | 0.0000 | 0.0006 | 0.0000 | 0.9010 | 0.2256 |
| | 10 : 1 | 0.0000 | 1.0000 | 0.0003 | 0.9888 | 0.8783 | 0.0040 | 1.0000 | 1.0000 | 1.0000 | 0.9989 | 0.0000 | 0.6246 |

| Trial4 | Test #: SPOP | Monte1 | Monte2 | Monte3 | Monte4 | Monte5 | Monte6 | Monte7 | Monte8 | Monte9 | Monte10 | Monte11 | Mean |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
|  | 1 : 0 | 1.0000 | 0.0167 | 1.0000 | 0.9963 | 1.0000 | 0.9630 | 0.8542 | 0.0000 | 0.0011 | 0.9922 | 1.0000 | 0.7112 |
| | 2 : 1 | 0.0000 | 0.0001 | 0.0000 | 0.0000 | 0.0086 | 0.9996 | 0.0001 | 0.0073 | 0.1754 | 0.0000 | 0.0001 | 0.1083 |
| | 3 : 0 | 1.0000 | 1.0000 | 1.0000 | 0.9972 | 0.0000 | 1.0000 | 0.9999 | 1.0000 | 1.0000 | 1.0000 | 0.0003 | 0.8179 |
| | 4 : 1 | 0.3789 | 1.0000 | 0.3241 | 0.0000 | 0.9964 | 1.0000 | 1.0000 | 0.9993 | 1.0000 | 0.9718 | 0.0078 | 0.6980 |
| | 5 : 0 | 0.0000 | 0.9928 | 0.9998 | 0.8345 | 0.0000 | 1.0000 | 0.9999 | 0.0000 | 0.9993 | 1.0000 | 0.9997 | 0.7115 |
| | 6 : 1 | 0.0039 | 1.0000 | 1.0000 | 0.9746 | 0.9988 | 0.9970 | 0.0001 | 0.0000 | 0.0038 | 1.0000 | 0.9988 | 0.6343 |
| | 7 : 0 | 0.0000 | 0.0065 | 0.0000 | 0.1120 | 0.0000 | 1.0000 | 0.9833 | 1.0000 | 0.0000 | 0.0000 | 0.0057 | 0.2825 |
| | 8 : 1 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 0.9983 | 1.0000 | 1.0000 | 1.0000 | 0.8101 | 1.0000 | 0.9926 | 0.9819 |
| | 9 : 0 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| | 10 : 1 | 0.1585 | 0.9483 | 0.2321 | 0.8237 | 0.9974 | 1.0000 | 1.0000 | 0.0000 | 1.0000 | 0.0134 | 0.0000 | 0.5612 |

| Trial5 | Test # : SPOP | Monte1 | Monte2 | Monte3 | Monte4 | Monte5 | Monte6 | Monte7 | Monte8 | Monte9 | Monte10 | Monte11 | Mean |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
|  | 1 : 0 | 0.0000 | 0.9732 | 1.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0001 | 0.0006 | 0.0000 | 0.0000 | 0.9951 | 0.2699 |
| | 2 : 1 | 0.9999 | 1.0000 | 1.0000 | 0.9998 | 0.0033 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 0.9094 |
| | 3 : 0 | 0.0000 | 0.0000 | 0.0033 | 0.0000 | 0.0000 | 1.0000 | 0.0000 | 0.9999 | 1.0000 | 1.0000 | 0.0007 | 0.3640 |
| | 4 : 1 | 0.0000 | 1.0000 | 0.9997 | 0.0068 | 0.0003 | 0.0008 | 0.4298 | 0.0000 | 0.0001 | 0.0000 | 0.0000 | 0.2216 |
| | 5 : 0 | 1.0000 | 0.0000 | 0.6263 | 0.0000 | 0.0000 | 0.9763 | 0.0000 | 0.0215 | 0.0002 | 0.5638 | 0.9029 | 0.3719 |
| | 6 : 1 | 1.0000 | 1.0000 | 0.9967 | 0.9997 | 1.0000 | 1.0000 | 0.0000 | 1.0000 | 0.6760 | 1.0000 | 1.0000 | 0.8793 |
| | 7 : 0 | 1.0000 | 1.0000 | 1.0000 | 0.9991 | 1.0000 | 0.9998 | 1.0000 | 0.0624 | 0.0100 | 1.0000 | 1.0000 | 0.8247 |
| | 8 : 1 | 0.0471 | 1.0000 | 1.0000 | 0.0008 | 0.0001 | 0.0052 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 0.9977 | 0.6410 |
| | 9 : 0 | 0.0442 | 0.0199 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0313 | 0.0000 | 0.0000 | 0.7992 | 0.0813 |
| | 10 : 1 | 0.0000 | 1.0000 | 0.4090 | 1.0000 | 0.0008 | 0.0129 | 0.0013 | 0.0000 | 0.0001 | 0.9782 | 1.0000 | 0.4002 |

| Trial6 | Test # : SPOP | Monte1 | Monte2 | Monte3 | Monte4 | Monte5 | Monte6 | Monte7 | Monte8 | Monte9 | Monte10 | Monte11 | Mean |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
|  | 1 : 0 | 0.0000 | 0.0755 | 0.0001 | 0.0001 | 0.0000 | 0.0002 | 0.9998 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0978 |
| | 2 : 1 | 1.0000 | 0.9999 | 0.0736 | 0.0494 | 0.0039 | 0.9867 | 1.0000 | 0.0003 | 0.9948 | 0.1567 | 0.0000 | 0.4787 |
| | 3 : 0 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 0.9729 | 1.0000 | 1.0000 | 1.0000 | 0.2159 | 1.0000 | 1.0000 | 0.9263 |
| | 4 : 1 | 0.8195 | 0.0009 | 0.2068 | 0.9985 | 0.0008 | 1.0000 | 1.0000 | 0.9998 | 0.9972 | 0.9983 | 0.9961 | 0.7289 |
| | 5 : 0 | 1.0000 | 1.0000 | 0.0000 | 1.0000 | 0.8761 | 1.0000 | 0.0061 | 0.0006 | 0.0000 | 0.9992 | 0.9335 | 0.6196 |
| | 6 : 1 | 1.0000 | 1.0000 | 0.9995 | 1.0000 | 0.0000 | 1.0000 | 0.9854 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 0.9077 |
| | 7 : 0 | 0.0000 | 0.0000 | 1.0000 | 0.3800 | 0.0000 | 0.0041 | 0.0000 | 0.0012 | 0.0000 | 0.0000 | 0.0000 | 0.1259 |
| | 8 : 1 | 0.0000 | 0.0179 | 0.9757 | 0.9564 | 0.0000 | 1.0000 | 0.0010 | 0.4151 | 1.0000 | 0.0003 | 0.0004 | 0.3970 |
| | 9 : 0 | 0.9971 | 0.2266 | 0.0038 | 0.4729 | 0.0000 | 1.0000 | 1.0000 | 0.1839 | 1.0000 | 0.9996 | 0.0002 | 0.5349 |
| | 10 : 1 | 0.0000 | 0.0000 | 0.9995 | 0.0000 | 0.0000 | 0.0000 | 0.0103 | 0.0150 | 0.9993 | 0.0000 | 0.9712 | 0.2723 |

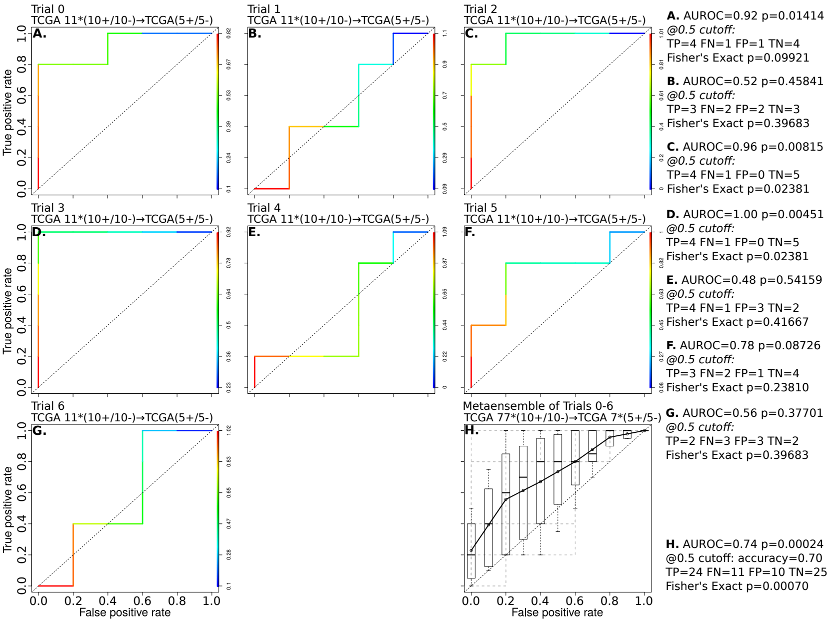


Figure S3.9: **Ensemble performance**, for comparison to ensemble performance in Fig 3.5C1. Classifiers in ensemble selected by highest validation accuracy. This estimates performance on unseen data. Performance for the classifiers instead selected by highest *test* set accuracy is shown in Fig S3.10.

Microscope slides were scanned at 0.25 ± 0.003 microns per pixel [$\mu pp$], using an Aperio AT2 scanner. The resulting SVS data file consists of multiple levels, where level 0 is not downsampled, level 1 is downsampled by a factor of 4, level 2 by a factor of 16, and level 3 by a factor of 32. From each level, 800x800px patches were extracted via the OpenSlide software library[42]. We refer to level 2 as low magnification and level 0 as high magnification. Level 2 approximately corresponds to a 10x eyepiece lens and 10x objective lens at the microscope when the scan is 0.5$\mu pp$. Our saliency predictor assumed 0.5$\mu pp$ scans, though the scans here were 0.25$\mu pp$, but appeared robust.

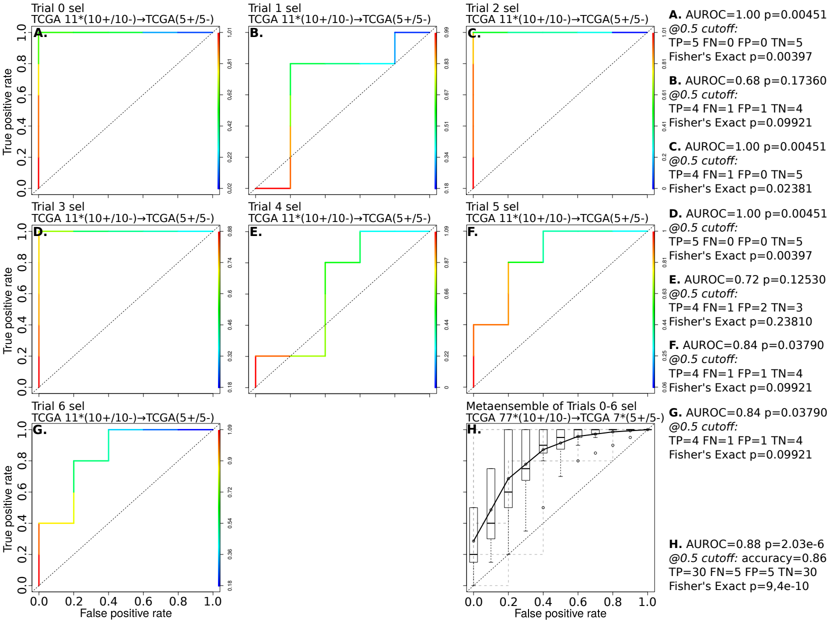


Figure S3.10: **Tuned ensemble performance**, for comparison to ensemble performance in Fig 3.5C. The tuned ensembles are used for prediction on the independent MSK-IMPACT cohort (Fig 3.5C2,C3). Tuning provides limited additional training for each ensemble via model selection on the corresponding TCGA test set. Tuned ensemble mutation prediction distributions shown in Fig S3.8.

Algorithm 3.1 describes data preprocessing for training. In prior work, we developed a patch saliency predictor[67]. A TMARKER classifier was trained to determine cell types[75]. We define the dominant tumor patch as having the maximum number of cancer cells of all 800x800px salient patches at low magnification. Within the 800x800px dominant tumor patch, we select an 800x800px patch at high magnification having the maximum number of abnormal cells. Malignant cells count 1 towards the maximum abnormal cell count, unknown cells count 0.5, and healthy cells count 0. The dominant tumor is explored in increments of 10 pixels, until the bounding box with the maximum number of abnormal cells is found. Whereas our saliency predictor operated on low-power tissue-level details of a patch, our SPOP predictor operates on high-power cell-level details of a patch. A patient is discarded from the study if the abnormal patch over-stained eosin blue, or is >50% blurred, or is >50% background, or is >50% blood – as determined by visual inspection.

Algorithm 3.2 describes the neural network ensemble for training. This procedure allows learning to occur despite a mere 20 patients having an SPOP mutation, compared to 157 not having an SPOP mutation. Additionally, having 5 ensembles – each trial yielding one ensemble of 11 ResNets – allows a CI in predicted SPOP mutation state to be calculated for both the generalization error estimate during training and any future patient (Algorithm 3.3). Moreover, such a large number of ResNets can fully sample the SPOP non-mutant patients, while each ResNet is still trained with an equal proportion of SPOP mutants and non-mutants. We use two neural network architectures, both the published ResNet-50 architecture ($r50$ in Alg 3.2) and our custom ResNet-50 with a 50% dropout[48] layer with an additional 1024 fully-connected layer as the top layer ($drp5$ in Alg 3.2, and shown as Drop50 in Fig 3.4). In practice, at least one architecture tended to have validation accuracy ≥ 0.6. Architecture diversity may increase intra-ensemble ResNet variance, and the decorrelation in errors should average out in the ensemble. Trials 0, 2, and 3 used two Drop50 learners and nine ResNet-50 classifiers (Fig 3.4, Table S3.1). Trial 1 had six and five, respectively. Trial 4 had three and eight. Trials 1 and 4 had the worst performance by AUROC (Fig 3.5), both had at least one ResNet-50 predictor with 0.3 or worse test set accuracy (Table S3.1)), and both had more than two Drop50 learners in the final ensemble for the trial. For challenging draws of training, validation, and test sets, Drop50 learners may slightly outperform ResNet-50 learners, though generalization accuracy may remain low due to the clustering of the data in the draws and limited sample sizes.

ResNets (both ResNet-50 and Drop50 architectures) were trained on 10 mutant and 10 non-mutant patients, so with data augmentation each ResNet was trained on 21600 images total. For each training set, a validation set of 5 mutant and 5 non-mutant patients was used and did not overlap with the training set, so with data augmentation each ResNet was validated against 10800 images. The test set for an ensemble consistent of 5 mutant and 5 non-mutant patients which were not augmented and did not overlap with any training or validation set for any classifier in the ensemble, so each ensemble was tested against 10 images.

There is remarkable variability in test accuracy among the testable models from each Monte Carlo cross validation run (Table S3.1). If the test set is drawn from approximately the same distribution as the validation set, where ResNet validation accuracy is 0.6+ and ResNets are uncorrelated, then we can expect 6 of the 11 ResNets ($6/11 < 0.6$) in an ensemble to correctly predict SPOP mutation state on average. In this way the ensemble is a strong learner based on ResNet weak learners[68,69]. Through ensemble averaging, the mean SPOP mutation probability is computed over all 11 constituent ResNets, to provide the final probability from the ensemble.

CHAPTER 4

# CONNECTING PATHOLOGISTS THROUGH PATIENT CASE SEARCH ON SOCIAL MEDIA AND PUBMED

Portions of this chapter first appeared in Schaumberg[*] et al 2020. Andrew J. Schaumberg originally wrote Schaumberg et al 2020, which was extensively edited by Mariam Aly. Thomas J. Fuchs and others reviewed Schaumberg et al 2020, including Laura G. Pastrián, Bobbi S. Pritt, Mario Prieto Pozuelo, Stephen Yip, Bin Xu, Aurélien Morini, Olaleke O. Folaranmi, Jerad M. Gardner, Celina Stayerman, and S. Joseph Sirintrapun. Schaumberg et al 2020 was a collaboration that additionally included Wendy C. Juarez-Nicanor, Sarah J. Choudhury, Ricardo Sotillo Sánchez, Khanh Ho, Nusrat Zahra, Betul Duygu Sener, Srinivas Rao Annavarapu, Karra A. Jones, Kathia Rosado-Orozco, Sanjay Mukhopadhyay, Carlos Miguel, Hongyu Yang, Rola H. Ali, Corina Rusu, Celina Stayerman, John Gross, and Dauda E. Suleiman. The author makes this chapter available under the terms of a Creative Commons Attribution-NonCommercial-NoDerivatives [CC-BY-NC-ND] 4.0 International license, at `https://creativecommons.org/licenses/by-nc-nd/4.0/`.

Early developments in the work documented in this chapter were presented as an extended abstract[16][1] and won Best Poster at Pathology Visions 2019. This work is currently *in press* at Modern Pathology[2].

---

[*] Schaumberg AJ, Juarez-Nicanor WC, Choudhury SJ, Pastrián LG, Pritt BS, Prieto-Pozuelo M, Sotillo-Sánchez R, Ho K, Zahra N, Sener BD, Yip S, Xu B, Annavarapu SR, Morini A, Jones KA, Rosado-Orozco K, Mukhopadhyay S, Miguel C, Yang H, Rosen Y, Ali RH, Folaranmi OO, Gardner JM, Rusu C, Stayerman C, Gross J, Suleiman DE, Sirintrapun SJ, Aly M, and Fuchs TJ. Interpretable multimodal deep learning for real-time pan-tissue pandisease pathology search on social media. *bioRxiv*, page 396663, March 2020. doi:10.1101/396663. Publisher: Cold Spring Harbor Laboratory Section

[1] Alternatively, this extended abstract is available at `http://www.jpathinformatics.org/downloadpdf.asp?issn=2153-3539;year=2020;volume=11;issue=1;spage=1;epage=1;aulast=;type=2`.

## 4.1 Introduction

The United Nations' Sustainable Development Goal 3: Good Health and Well-Being suggests that it is essential to "ensure healthy lives and promote well-being for all at all ages"[2]. In the furtherance of this goal, it is suggested to "[s]ubstantially increase [...] the recruitment, development, training, and retention of the health workforce in developing countries" to universally achieve "access to quality essential health care services"[2]. We therefore take connecting pathologists worldwide to be important.

Indeed, Nix *et al*.[76] find pathologists in developing countries (e.g. India, Brazil, and Pakistan) frequently use social media, and 220/1014 (22%) of the posts they analyzed involved "asking for opinions on diagnosis". The use of social media by pathologists occurs worldwide for both challenging cases and education[77,78,79]. This suggests social media can facilitate global collaborations among pathologists for novel discoveries[80]. We expand on these approaches by combining (i) real-time machine learning with (ii) expert pathologist opinions via social media to facilitate (i) search for similar cases and (ii) pathological diagnosis by sharing expertise on a particular disease, often with underserved hospitals.

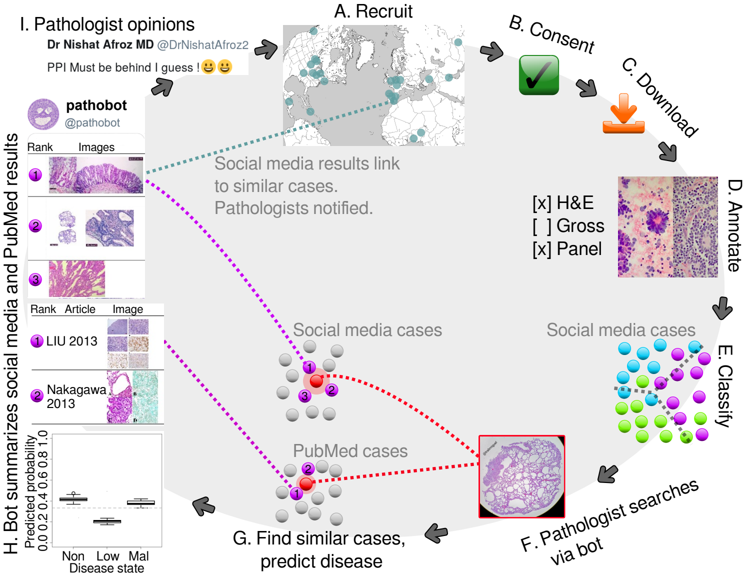
For machine learning to work in general practice, it must be trained on data (i) of sufficient diversity to represent the true variability of what is observed (ii) in a sufficiently realistic setting that may differ from tightly controlled experimental conditions[81]. We therefore (i) collaborate with pathologists worldwide where we (ii) use for training the images that these pathologists share to obtain opinions, which are often histopathology microscopy pictures from a smartphone. We did not observe many images from whole slide scanners, which at a global scale have been adopted slowly, due in part to cost and complexities of digital pathology workflows[82,83].

---

[2] The online version of the Schaumberg* et al 2020 article in Modern Pathology is expected to be available at `https://doi.org/10.1038/s41379-020-0540-1`.

For machine learning to work accurately, it must be trained on a sufficiently large dataset. Our first aim is therefore to curate a large dataset of pathology images for training a machine learning classifier. This is important because in other machine learning domains, e.g. natural vision tasks, datasets of millions of images are often used to train and benchmark, e.g. ImageNet[51] or CIFAR-10[84]. Transfer learning allows limited repurposing of these classifiers for other domains, e.g. pathology[11,85,86,87]. Indeed, we[5] are among many who start in computational pathology[25] with deep-neural networks pre-trained on ImageNet[88,89,90], and we do so here.

However, computational pathology datasets annotated for supervised learning are often much smaller than millions of images. For example, there are only 32 cases in the training data for a Medical Image Computing and Computer Assisted Intervention challenge (available at `http://miccai.cloudapp.net/competitions/82`) for distinguishing brain cancer subtypes, and this includes both pathology and radiology images. Other studies are larger, such as the TUmor Proliferation Assessment Challenge (TUPAC16) dataset of 821 cases[91] – all 821 cases being whole slide images from The Cancer Genome Atlas (TCGA) (`http://cancergenome.nih.gov/`). TCGA has tens of thousands of whole slide images available in total, but these images are only hematoxylin


I. Pathologist opinions
A. Recruit
B. Consent
C. Download
D. Annotate
[x] H&E
[ ] Gross
[x] Panel
E. Classify
F. Pathologist searches via bot
G. Find similar cases, predict disease
H. Bot summarizes social media and PubMed results
Social media results link to similar cases. Pathologists notified.
Social media cases
Social media cases
PubMed cases
pathobot
@pathobot
Rank
Images
Rank
Article
Image
LIU 2013
Nakagawa 2013
Predicted probability
0.0 0.2 0.4 0.6 0.8 1.0
Non
Low
Mal
Disease state

**Dr Nishat Afroz MD** @DrNishatAfroz2

PPI Must be behind I guess !😀😀

Figure 4.1: **Graphical summary.** Pathologists are recruited worldwide **(A)**. If a pathologist consents to having their images used **(B)**, we download those images **(C)** and manually annotate them **(D)**. Next, we train a Random Forest classifier to predict image characteristics, e.g. disease state **(E)**. This classifier is used to predict disease and search. If a pathologist posts a case to social media and mentions @pathobot **(F)**, our bot will use the post's text and images to find similar cases on social media and PubMed **(G)**. The bot then posts summaries and notifies pathologists with similar cases **(H)**. Pathologists discuss the results **(I)**, and some also decide to share their cases with us, initiating the cycle again **(A)**. *Procedure overview* in the supplement explains further (Sec S4.5.4).

and eosin (H&E) stained slides, and do not represent non-neoplastic lesions such as infections, which are clinically important to correctly diagnose[92]. The main limitation is that obtaining annotations from a pathologist is difficult due to outstanding clinical service obligations, which prevented our earlier efforts from scaling up[67]. We overcome this limitation by curating a large and diverse dataset of 13,626 images from Twitter and 113,161 images from PubMed, where text annotations came from social media post text, hashtags, article titles, abstracts, and/or figure captions.

Equipped with our large dataset, we then address our second main aim, which is to utilize machine learning trained on this dataset to facilitate prospective disease state predictions and search from pathologists in real-time on social media. To that end, we capitalize on a common and systematic approach to diagnosis in which a disease is in one of three classes[92]. Specifically, we use machine learning on pathology images from social media and PubMed to classify images into one of three disease states: *nontumor* (e.g. normal, artifact (Fig S4.11), injury, infection, or nontumor), *low grade* (e.g. pre-neoplastic, benign, or low grade malignant potential), or *malignant*.

We then implement a social media bot that in real time applies our machine learning classifiers in response to pathologists on social media to (i) search for similar cases, (ii) provide quantitative predictions of disease states, and (iii) encourage discussion (Fig 4.1). When this bot links to a similar case, the pathologist who shared that case is notified. The ensuing discussions among pathologists are more informative and context-specific than a computational prediction. For instance, to make a diagnosis of Kaposi's sarcoma, first-world countries have access to an HHV8 histopathology stain, but a pathologist in a developing country may instead be advised to check patient history of HIV because the HHV8 stain is prohibitively expensive. Obviously, a computational prediction of cancer/non-cancer is far less helpful than what humans do: discuss.

In order for machine learning approaches to be useful in a clinical setting, it is critical that these approaches be interpretable and undergo rigorous prospective testing[93]. Furthermore, these approaches need to be accompanied by quantified measures of prediction uncertainty[94]. It may be argued whenever human life is at risk – (i) interpretability, (ii) uncertainty quantification, and (iii) prospective testing are essential – whether the context is medicine or self-driving cars[95,96]. Our social media bot and methods are the first in computational pathology to meet all of these criteria in that (i) we provide multiple levels of interpretability (e.g. Random Forest feature importance and deep learning activation heatmaps), (ii) we statistically quantify prediction uncertainty using ensemble methods, and (iii) we prospectively test in full public view on social media. Concretely, this means (i) a pathologist can interpret what concepts the machine learning finds to be diagnostic in general or what parts of a particular image suggest a specific disease state, (ii) statistical significance, confidence intervals, or boxplots of computational predictions are presented to a pathologist for assessment (e.g. the boxplot in Fig 4.1 *lower left*), and (iii) in real time a pathologist can interact with our social media bot and method to appraise performance on a case-by-case basis, as well as evaluate the public history of pathologist-bot interactions on social media.

## 4.2 Materials and methods

This study was approved by the Institutional Review Board at Memorial Sloan Kettering Cancer Center.

### 4.2.1 Social media data

From Twitter we curate 13,626 images from 6,351 tweets from 25 pathologists from 13 countries. We chose Twitter primarily for its brevity, i.e. one Tweet is

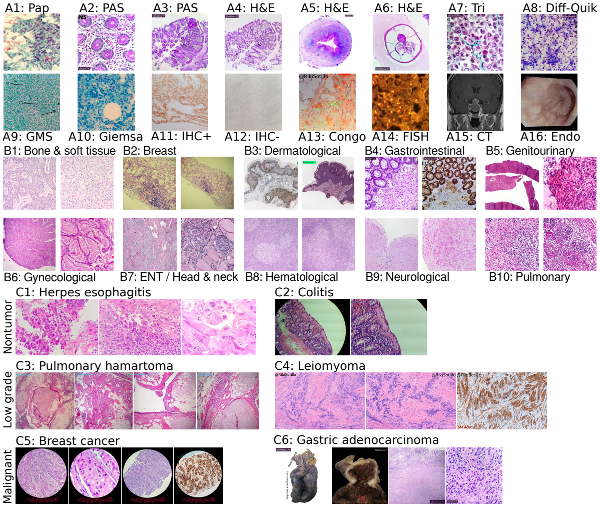

A1: Pap
A2: PAS
A3: PAS
A4: H&E
A5: H&E
A6: H&E
A7: Tri
A8: Diff-Quik
A9: GMS
A10: Giemsa
A11: IHC+
A12: IHC-
A13: Congo
A14: FISH
A15: CT
A16: Endo
B1: Bone & soft tissue
B2: Breast
B3: Dermatological
B4: Gastrointestinal
B5: Genitourinary
B6: Gynecological
B7: ENT / Head & neck
B8: Hematological
B9: Neurological
B10: Pulmonary
Nontumor
C1: Herpes esophagitis
C2: Colitis
Low grade
C3: Pulmonary hamartoma
C4: Leiomyoma
Malignant
C5: Breast cancer
C6: Gastric adenocarcinoma

Figure 4.2: **Technique, tissue, and disease diversity.** Panel set A shows diverse techniques in our data. Initials indicate author owning image. **(A1)** *R.S.S.*: Papanicolaou stain. **(A2)** *L.G.P.*: Periodic acid-Schiff (PAS) stain, glycogen in pink. **(A3)** *L.G.P.*: PAS stain, lower magnification. **(A4)** *L.G.P.*: H&E stain c.f. Panel A3. **(A5)** *L.G.P.*: H&E stain, human appendix, including parasite *Enterobius vermicularis* (c.f. Fig S4.12). **(A6)** *L.G.P.*: Higher magnification *E. vermicularis* c.f. Panel A5. **(A7)** *L.G.P.*: Gömöri trichrome, collagen in green. **(A8)** *L.G.P.*: Diff-quik stain, for cytology. **(A9)** *R.S.S.*: GMS stain (*Intra-stain diversity* in supplement details variants, Sec S4.5.3), fungi black. **(A10)** *M.P.P.*: Giemsa stain. **(A11)** *A.M.*: Immunohistochemistry (IHC) stain, positive result. **(A12)** *A.M.*: IHC stain, negative result. **(A13)** *R.S.S.*: Congo red, polarized light, plaques showing green birefringence. **(A14)** *M.P.P.*: Fluorescence *in situ* hybridization (FISH) indicating breast cancer *Her2* heterogeneity. **(A15)** *S.Y.*: Head computed tomography (CT) scan. **(A16)** *L.G.P.*: Esophageal endoscopy.

In panel set B we show differing morphologies for all ten histopathological tissue types on Twitter. **(B1)** *C.S.*: bone and soft tissue. We include cardiac here. **(B2)** *K.H.*: breast. **(B3)** *R.S.S.*: dermatological. **(B4)** *L.G.P.*: gastrointestinal. **(B5)** *O.O.F.*: genitourinary. **(B6)** *M.P.P.*: gynecological. **(B7)** *B.X.*: otorhinolaryngological a.k.a. head and neck. We include ocular, oral, and endocrine here. **(B8)** *C.S.*: hematological, e.g. lymph node. **(B9)** *S.Y.*: neurological. **(B10)** *S.M.*: pulmonary.

In panel set C we show the three disease states we use: nontumor, low grade, and malignant. **(C1)** *M.P.P.*: Nontumor disease, i.e. herpes esophagitis with Cowdry A inclusions. **(C2)** *K.H.*: Nontumor disease, i.e. collagenous colitis showing thickened irregular subepithelial collagen table with entrapped fibroblasts, vessels, and inflammatory cells. **(C3)** *A.M.*: Low grade, i.e. pulmonary hamartoma showing entrapped clefts lined by respiratory epithelium. **(C4)** *R.S.S.*: Low grade, i.e. leiomyoma showing nuclear palisading. We show IHC completeness but it is not included for machine learning. **(C5)** *B.D.S.*: Malignant, i.e. breast cancer with apocrine differentiation. **(C6)** *L.G.P.*: Malignant, i.e. relapsed gastric adenocarcinoma with diffuse growth throughout the anastomosis and colon. Gross sections (e.g. Fig S4.13) shown for completeness but not used.

at most 280 characters, so we did not expect to need complicated text processing logic to parse tissues or diagnoses. Written permission to download and use the data was obtained from each collaborating pathologist. One pathologist publicly declared their data free to use, so we use these data with acknowledgement. One pathologist donated his glass slide library to another pathologist, and the receiving pathologist shared some received cases on social media, which we treat as belonging to the receiving pathologist. Images are annotated with their tweet text and replies. We use these data for supervised learning.

### 4.2.2 PubMed data

To represent PubMed data, we download the PubMed Central "Open Access Subset" of 1,074,484 articles. We first trained a classifier to distinguish H&E images from all others on social media (Figs 4.2, S4.14, S4.15), then used the classifier to identify PubMed articles that have at least one H&E figure. From the identified 30,585 articles we retain 113,161 H&E images to comprise our PubMed dataset. Images are annotated with figure caption, article abstract, and article title. This expanded dataset may contain disease that is too rare to be represented in social media data.

### 4.2.3 Image processing

We manually curate all social media images, separating pathology from non-pathology images. *Defining an acceptable pathology image* (Sec S4.5.1) details this distinction in the supplement (Fig S4.14). Some pathologists use our Integrated

Pathology Annotator (IPA) tool to browse their data and manually curate the annotations for their cases (Figs S4.16, S4.17). We retain non-pathology data publicly posted by consenting pathologists that cannot be publicly distributed to enable building a machine learning classifier that can reliably distinguish pathology from non-pathology images.

### 4.2.4 Text processing

*Text data overview* (Sec S4.5.5) in the supplement discusses our text processing to derive ground truth from social media posts (Fig S4.18). We use hashtags, e.g. #dermpath and #cancer, as labels for supervised learning. We process the text of the tweet and the replies, detecting terms that indicate tissue type or disease state. For instance, "ovarian" typically indicates gynecological pathology, while "carcinoma *in situ*" typically indicates low grade disease (specifically, preneoplastic disease in our low grade disease state category). Our text processing algorithm (Fig S4.18) is the result of author consensus.

### 4.2.5 Random Forest classifier

We train a Random Forest of 1,000 trees as a baseline for all tasks. A white-balanced image is scaled so its shortest dimension is 512 pixels (px). White balancing helps correct images with reduced blue coloration due to low lighting (Fig S4.15D). The 512×512px center crop is then extracted, and 2,412 hand-engineered image features are calculated for this crop (Figs 4.3, S4.19).

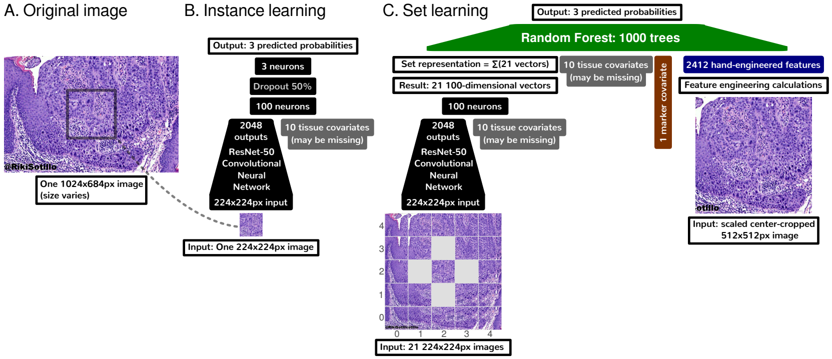

A. Original image
B. Instance learning
C. Set learning
@RikiSotillo
One 1024x684px image (size varies)
Output: 3 predicted probabilities
3 neurons
Dropout 50%
100 neurons
2048 outputs
ResNet-50 Convolutional Neural Network
224x224px input
10 tissue covariates (may be missing)
Input: One 224x224px image
Output: 3 predicted probabilities
Random Forest: 1000 trees
Set representation = Σ(21 vectors)
Result: 21 100-dimensional vectors
10 tissue covariates (may be missing)
1 marker covariate
2412 hand-engineered features
Feature engineering calculations
100 neurons
2048 outputs
ResNet-50 Convolutional Neural Network
224x224px input
10 tissue covariates (may be missing)
4
3
2
1
0
0
1
2
3
4
Input: 21 224x224px images
Input: scaled center-cropped 512x512px image

Figure 4.3: **Deep learning methods summary. (A)** An overall input image may be of any size, but must be at least 512×512 pixels (px). **(B)** We use a ResNet-50[62] deep convolutional neural network to learn to predict disease state (nontumor, low grade, or malignant) on the basis of a small 224×224px patch. This small size is required to fit the ResNet-50 and image batches in limited GPU memory. **(C)** For set learning, this network transforms each of the 21 patches sampled evenly from the image in a grid to a 100-dimensional vector. These 21 patches span the overall input image entirely. For instance, if the overall input image is especially wide, the 21 patches will overlap less in the X dimension. The ResNet-50 converts these 21 patches to 21 vectors. These 21 vectors are summed to represent the overall image, regardless of the original image's size, which may vary. To represent additional clinico-visual context of a patient case, this sum vector is concatenated with tissue covariates (which may be missing for some images), marker mention covariate, and hand-engineered features. A Random Forest then learns to predict disease state on this concatenation that encodes (i) task-agnostic hand-engineered features (Fig S4.19) near the image center, (ii) task-specific features from deep learning throughout the image, (iii) whether IHC or other markers were mentioned for this case, and (iv) optionally tissue type. Other machine learning tasks, e.g. histology stain prediction and tissue type prediction, were simpler. For simpler tasks, we used only the Random Forest and 2,412 hand-engineered features, without deep learning.

### 4.2.6 Customized hybrid deep-learning-random-forest model and clinical covariates

**Image preprocessing and data augmentation** For image preprocessing, a white-balanced image is scaled to be 512 pixels in its shortest dimension, and for deep learning, 224×224px patches are sampled to train a deep convolutional neural network. For deep learning, we use data augmentation of random rotations, random flips, random zoom/rescaling, random brightness variations, Gaussian noise, and Mixup[97]. This means that throughout training hundreds of times over our data we make many small changes to the data each time, e.g. to teach the neural network that rotating an image does not change the diagnosis. *Deep learning* (Sec S4.5.11) discusses further.

**Deep learning and deep features** To maximize performance by learning disease-state-specific features, we additionally consider deep learning for the most challenging task of disease state prediction. Our deep learning architecture is a ResNet-50[62] (Fig 4.3B) pretrained on ImageNet, which we train end-to-end without freezing layers (Fig S4.23). This means the ResNet-50 deep convolutional neural network is initially trained to classify natural images, e.g. cats and dogs, but every neuron may be adjusted in a data-driven manner for histology-specific learning on our pathology Twitter dataset. To determine how deep feature representations change before and after training the ResNet-50 on histopathology images and covariates, we analyze both (i) $\text{ImageNet}_{2048}$ features from the ResNet-50 that has *not* been trained on histopathology data, and (ii) 100 deep features based on the same ResNet-50 where all neurons *have* been further trained on histopathology data. We define $\text{ImageNet}_{2048}$ features as the 2,048 outputs from the ResNet-50's final Global Average Pooling layer, summed over 21 image patches in a grid fashion and concatenated with other features for Random Forest learning (Fig 4.3C). For histopathology deep learning, we append a 100-neuron fully-connected layer atop the ResNet-50, connecting to the ResNet-50 and covariates, and sum over the same 21 image patches in a grid fashion (Fig 4.3B). *Deep learning instance and set feature vectors* (Sec S4.5.8) discusses this and the feature interpretability related to the Heaviside step function (Eqns 4.6, 4.8).

**Clinical covariates** To best predict disease state and find similar cases, we seek to include as much patient-related context as possible in our computational pathology machine learning models, so we additionally include clinical information, i.e. tissue type and marker mentions. To represent the tissue type covariate, we include a ten-dimensional one-hot-encoded binary vector to encode which one of the ten possible tissue types is present for this case. If the tissue type is unknown, tissue type is all zeroes for the neural network while being missing values for the Random Forest. We also include a binary one-dimensional marker mention covariate, which is 1 if any pathologist discussing the case mentions a marker test, e.g. “IHC” or “desmin”.

### 4.2.7 Disease state classifier repurposed for similarity-based search

After we train a *Random Forest classifier* (Sec 4.2.5) to predict/classify disease state from a variety of deep and non-deep features (Fig 4.3C), we then use this classifier’s Random Forest similarity metric for search[70,98]. Specifically, our Random Forest consists of 1,000 Random Trees, each of which predicts disease state. If any given Random Tree makes an identical sequence of decisions to classify two histopathology images (each with optionally associated clinical covariates), the similarity of those two images is incremented by one. Aggregating across all Random Trees, the similarity of any two images can therefore be quantified as a number between 0 (not similar according to any Random Tree) and 1,000 (similar according to all 1,000 Random Trees). Equipped with this similarity metric, we repurpose the classifier for search: the classifier takes in a search image and compares it to each other image using this similarity metric, then provides a list of images ranked by similarity to the search image. This approach provides the first pan-tissue (i.e. bone and soft tissue, breast, dermatological, gastrointestinal, genitourinary, gynecological, head and neck, hematological, neurological, pulmonary, etc) pan-disease (i.e. nontumor, low grade, and malignant) patient case search in pathology.

### 4.2.8 Three levels of sanity checking for search

To inform the physician and to avoid mistakes, sanity checks are important in medicine, or wherever human life may be at risk. Quantifying uncertainty is particularly important[94] in medicine, to assess how much trust to put in predictions that will affect the patient's care. We are the first to offer three sanity checks for each individual search: (i) prediction uncertainty, (ii) prediction as a check for search, and (iii) prediction heatmaps. *Machine learning sanity checking for search* discusses further (Sec S4.5.9). Briefly, "prediction uncertainty" relies on an ensemble/collection of classifiers to assess if disease state prediction strength is statistically significant, and if not, the prediction and search using this image should not be trusted. Second, "prediction as a check for search" indicates that if the disease state classification for a given image is assessed as incorrect by a pathologist, search results using this image should not be trusted, because the same classifier is used for both prediction and search. Third, we use "prediction heatmaps" to show disease-state predictions for each subregion of a given image, based on deep learning. If a pathologist disagrees with these heatmaps, deep-learning-based search for that image cannot be trusted. A failure of any one of these three checks indicates that search results may be incorrect, and they are flagged as such.

### 4.2.9 Five levels of method interpretability

Interpretability is critical in medicine[93] for physicians to understand whether or not the machine learning is misinterpreting the data. For example, machine learning may uncover that pneumonia patients with a history of asthma have lower mortality risk, suggesting that asthma is protective against pneumonia mortality. However, this would not make sense to a physician, who would instead realize that such patients have lower mortality because they are more likely to be admitted directly to an intensive care unit[99,100]. Asthma is not protective from pneumonia mortality, intensive care is.

Ideally, interpretability facilitates both deductive and inductive human reasoning about the machine learning findings. Deductively, interpretability allows human reasoning about what machine learning finds in specific patient cases, e.g. explaining the malignant prediction overall for a patient by spatially localizing where malignancy-related features are in a histology image. Inductively, interpretability allows human reasoning about broad principles that may be inferred from the machine learning findings overall for a task, e.g. texture features are important in disease state prediction. To the best of our knowledge, it is novel to offer both deductive and inductive interpretability in a pan-tissue pan-disease manner in computational pathology. We do this with (i)

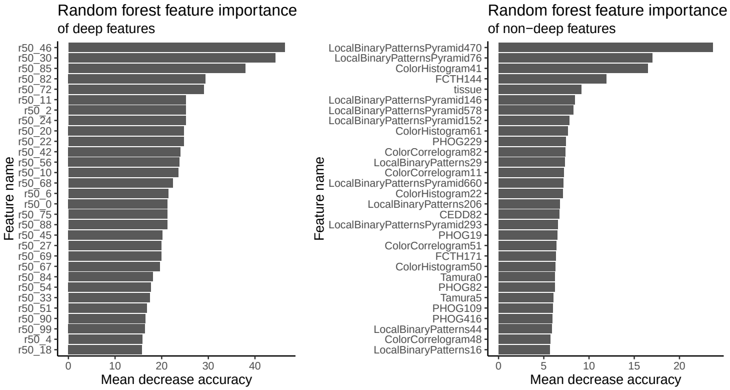


Figure 4.4: **Random Forest feature importance for prioritizing deep features, when non-deep, deep, and clinical features are used together for learning.** We use the mean decrease in accuracy to measure Random Forest feature importance. To do this, first, a Random Forest is trained on task-agnostic hand-engineered features (e.g. color histograms), task-specific deep features (i.e. from the ResNet-50), and the tissue type covariate that may be missing for some patients. Second, to measure the importance of a feature, we randomly permute/shuffle the feature's values, then report the Random Forest's decrease in accuracy. When shuffling a feature's values this way, more important features result in a greater decrease in accuracy, because accurate prediction relies on these features more. We show the most important features at the top of these plots, in decreasing order of importance, for deep features (*at left*) and non-deep features (*at right*). The most important deep feature is "r50_46", which is the output of neuron 47 of 100 (first neuron is 0, last is 99), in the 100-neuron layer we append to the ResNet-50 and train on histopathology images. Thus of all 100 deep features, r50_46 may be prioritized first for interpretation. Of non-deep features, the most important features include Local Binary Patterns Pyramid (LBPP), color histograms, and "tissue" (the tissue type covariate). LBPP and color histograms are visual features, while tissue type is a clinical covariate. LBPP are pyramid-based grayscale texture features that are scale-invariant and color-invariant. LBPP features may be important because we neither control the magnification a pathologist uses for a pathology photo, nor do we control staining protocol. For a before-and-after-training comparison that may suggest the histopathology-trained deep features represent edges, colors, and tissue type rather than texture, we also analyze feature importance of only-natural-image-trained ImageNet$_{2048}$ deep features in conjunction with hand-engineered features (Fig S4.20). *Marker mention and SIFT features excluded from Random Forest feature importance analysis* discusses other details in the supplement (Sec S4.5.10).

hand-engineered feature interpretability (Fig S4.19), (ii) Random Forest feature importance (Fig 4.4), (iii) before-and-after-histopathology-training feature importance comparison of deep features to hand-engineered features (Fig 4.4 vs Fig S4.20), (iv) deep feature activation maps (Figs 4.5D, S4.21), and (v) cluster analyses (Figs 4.6). *Machine learning interpretability for search* in the supplement discusses further (Sec S4.5.10).

**Histopathology-trained deep features represent edges, colors, and tissue** To understand what deep features learn to represent after training on histopathology data, we compare Random Forest feature importances of (a) $\text{ImageNet}_{2048}$ deep features [not trained on histopathology data] with hand-engineered features and tissue covariate (Fig S4.20), to (b) 100 deep features [trained on histopathology data] with hand-engineered features and tissue covariate (Fig 4.4). Before the deep neural network is trained on histopathology data, the tissue covariate as well as edge and color hand-engineered features are important (Fig S4.20). However, after the deep neural network is trained on histopathology data, tissue is less important while texture hand-engineered features are more important (Fig 4.4). Therefore, we reason that the deep neural network learns histopathology-relevant edge, color, and tissue features from histopathology data (which reduces the importance of e.g. hand-engineered edge and color features after learning), but the deep neural network may forget histopathology-relevant texture features during learning (which increases the importance of hand-engineered texture features after learning).

**Interpretability uncovers spatial prediction-to-feature correspondences of disease** Considering both introspective/inductive interpretability (Fig 4.4)

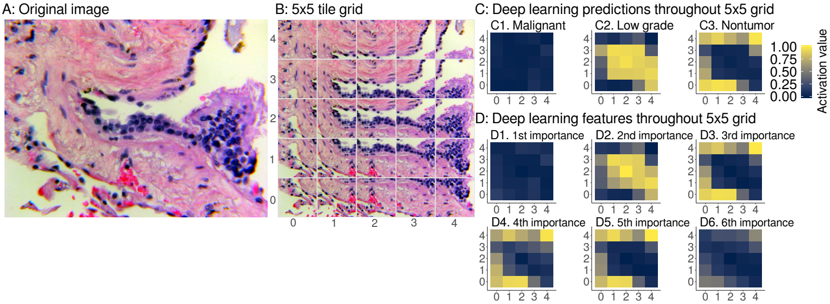

A: Original image
B: 5x5 tile grid
C: Deep learning predictions throughout 5x5 grid
C1. Malignant
C2. Low grade
C3. Nontumor
1.00
0.75
0.50
0.25
0.00
Activation value
D: Deep learning features throughout 5x5 grid
D1. 1st importance
D2. 2nd importance
D3. 3rd importance
D4. 4th importance
D5. 5th importance
D6. 6th importance
0 1 2 3 4

Figure 4.5: **Interpretable spatial distribution of deep learning predictions and features. (A)** An example image for deep learning prediction interpretation, specifically a pulmonary vein lined by enlarged hyperplastic cells, which we consider to be low grade disease state. Case provided by *Y.R.* **(B)** The image is tiled into a 5×5 grid of overlapping 224×224px image patches. For heatmaps, we use the same 5×5 grid as in Fig 4.1C *bottom left*, imputing with the median of the four nearest neighbors for 4 of 25 grid tiles. **(C)** We show deep learning predictions for disease state of image patches. **(C1)** throughout the image, predictions have a weak activation value of ˜0 for malignant, so these patches are not predicted to be malignant. **(C2)** the centermost patches have a strong activation value of ˜1, so these patches are predicted to be low grade. This spatial localization highlights the hyperplastic cells as low grade. **(C3)** the remaining normal tissue and background patches are predicted to be nontumor disease state. Due to our use of softmax, we note that the sum of malignant, low grade, and nontumor prediction activation values for a patch equals 1, like probabilities sum to 1, but our predictions are not Gaussian-distributed probabilities. **(D)** We apply the same heatmap approach to interpret our ResNet-50 deep features as well. **(D1)** the most important deep feature corresponds to the the majority class prediction, i.e. C1, malignant. **(D2)** the second most important deep feature corresponds to prediction of the second most abundant class, i.e. C2, low grade. **(D3)** the third most important deep feature corresponds to prediction of the third most abundant class, i.e. C3, nontumor. The fourth **(D4)** and fifth **(D5)** most important features also correspond to nontumor. **(D6)** the sixth most important deep feature does not have a clear correspondence when we interpret the deep learning for this case and other cases (Fig S4.21), so we stop interpretation here. As expected, we did not find ImageNet$_{2048}$ features to be interpretable from heatmaps, because these are not trained on histpathology (Fig S4.21A5).

and demonstrative/deductive interpretability (Fig 4.5), we find a correspondence between important deep features (Fig 4.4) and the spatial localization of deep learning predictions of disease state (Fig 4.5). Moreover, we find that using (Eqn 4.14) the three most important interpretable deep features slightly but significantly improve search performance (Table S4.1). *Deep set learning feature interpretation* discusses further (Sec S4.5.11).

**Deep features trained on histopathology logically cluster patients by disease state, whereas pathology-agnostic features do not** Through cluster analysis we interpret which features (i.e. hand-engineered, only-natural-image-trained, or histopathology-trained), if any, separate patients into meaningful groups, and if the features "make sense" to describe patient histopathology. As expected, neither hand-engineered features (Fig 4.6A1) nor only-natural-image-trained ImageNet$_{2048}$ deep features (Fig 4.6B1) cluster patient cases by disease state, presumably because these features are not based on histopathology. These approaches also do not cluster patients by contributing pathologist (Fig 4.6A2,B2) or by tissue type (Fig 4.6A3,B3). Additionally, we do not find that reducing dimensionality through principal components analysis qualitatively changes the clustering (Fig S4.22). In contrast, deep features trained on histopathology data do cluster patients together by disease state (Fig 4.6C1), but not by pathologist (Fig 4.6C2) or tissue (Fig 4.6C3). We conclude that these deep features primarily reflect representations of disease state in a non-tissue-specific manner. It is important to note that any clustering-based result must be carefully scrutinized, because features may suffer from artifacts, e.g. which pathologist shared the patient case. If taken to an extreme, learning to predict disease state on the basis of pathologist-specific staining/lighting/camera artifacts amounts to learning concepts such as, “if pathologist X typically shares images of malignant cases, and a new image appears to be from pathologist X, then this image probably shows malignancy”, which does not “make sense” as a way to predict disease state. Although we did not observe robust clustering by pathologist, even vague grouping by pathologist (Fig 4.6A2 *at gray arrows*) highlights the importance of critically assessing results. Artifact learning risk is one reason why we (i) rigorously test search through leave-one-pathologist-out cross-validation, and (ii) provide sanity checks.

### 4.2.10 Experimental design and evaluation

We evaluate our classifiers using 10-fold cross-validation to estimate bounds of accuracy and Area Under Receiver Operating Characteristic (AUROC) performance metrics. *Supplementary experimental design and evaluation* explains further (Sec S4.5.14). Because we intend for our methods to accurately find similar cases for any pathologist worldwide, we rigorously test search using leave-one-pathologist-out cross-validation and report precision@k. Leave-one-pathologist-out cross-validation isolates pathologist cases from one another, so a test set is independent from the corresponding training set. This isolates to a test set pathologist-specific or institution-specific imaging artifacts that may occur from microscopy, lighting, camera, or staining protocol. Thus our leave-one- pathologist-out cross-validation measurements quantify our method’s reproducibility, which is critical to measure in medical machine learning[81].

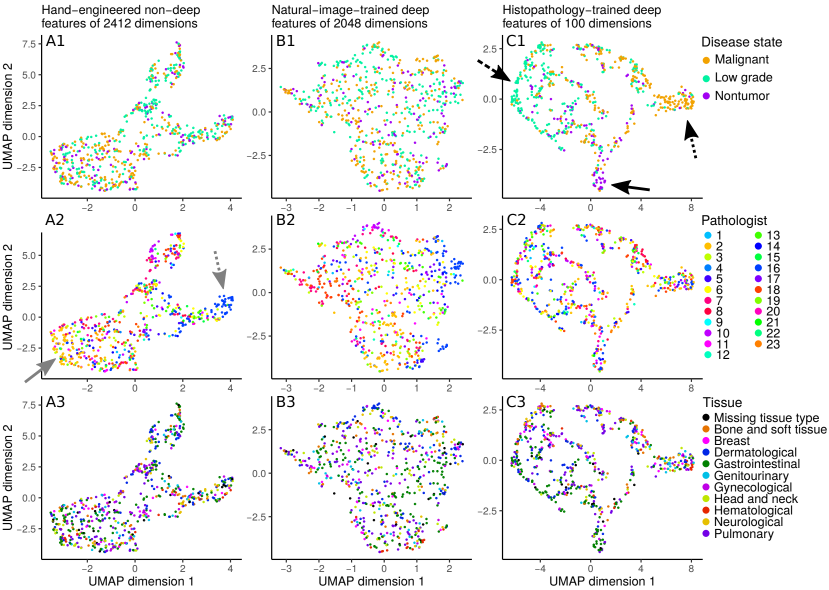


Figure 4.6: **Disease state clusters based on hand-engineered, natural-image-trained deep features, or histopathology-trained deep features.** To determine which features meaningfully group patients together, we apply the UMAP[101] clustering algorithm on a held-out set of 10% of our disease state data. Each dot represents an image from a patient case. In general, two dots close together means these two images have similar features. Columns indicate the features used for clustering: hand-engineered features (*at left column*), only-image-trained ImageNet$_{2048}$ deep features (*at middle column*), or histopathology-trained deep features (*at right column*). Rows indicate how dots are colored: by disease state (*at top row*), by contributing pathologist (*at middle row*), or by tissue type (*at bottom row*). For hand-engineered features, regardless of whether patient cases are labeled by disease state **(A1)**, pathologist **(A2)**, or tissue type **(A3)**, there is no strong clustering of like-labeled cases. Similarly, for only-natural-image-trained ImageNet$_{2048}$ deep features, there is no obvious clustering by disease state **(B1)**, pathologist **(B2)**, or tissue type **(B3)**. However, for histopathology-trained deep features , patient cases cluster by disease state **(C1)**, with separation of malignant (*at dotted arrow*), low grade (*at dashed arrow*), and nontumor (*at solid arrow*). There is no clear clustering by pathologist **(C2)** or tissue type **(C3)**. The main text notes that hand-engineered features may vaguely group by pathologist (**A2**, pathologists 2 and 16 *at solid and dotted arrows*).

### 4.2.11 Social media bot for public prospective testing

We present the first pathology-specific social media bot, @pathobot, on Twitter. This bot is a case similarity search tool that applies our methods. Pathologists on Twitter mention the bot in a tweet containing an image. The bot uses our Random Forest classifier to provide disease-state prediction for that image, and search for similar results. Its prediction and search results, along with quantitative assessments of prediction uncertainty, are provided to pathologists in real time. In this way, the bot facilitates prospective tests, and encourages collaboration: as pathologists use the bot, they provide us with complementary qualitative feedback and help us recruit additional collaborators. In this way, the bot facilitates prospective tests, and encourages collaboration: as pathologists publicly use the bot, they provide us with complementary qualitative feedback and these interactions help us recruit additional collaborators.

### 4.2.12 Computational hardware

For machine learning, we use Weka version 3.8.1[102] on a laptop with OpenJDK Java version 1.8.0_131. For deep learning, we use Tensorflow version 1.0.0 with Keras version 2.1.4[103] on a supercomputing cluster having GPUs supporting nVidia CUDA version 8.0 and cuDNN version 5.1. *Supplemetary computational hardware and software* discusses further (Sec S4.5.15). In R, we perform feature importance analyses with the randomForest package[104] and cluster analyses with the umap package[105].

## 4.3 Results

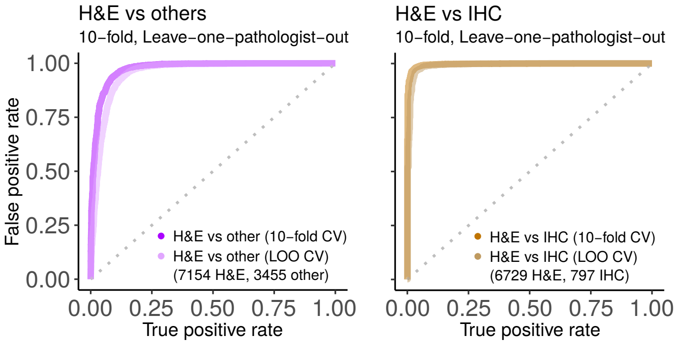


Figure 4.7: **H&E performance.** Predicting if an image is acceptable H&E human tissue or not (*at left*), or if image is H&E rather than IHC (*at right*). Ten replicates of tenfold cross-validation (10-fold) and leave-one-pathologist-out cross-validation (LOO) had similarly strong performance. This suggests the classifier may generalize well to other datasets. We use the "H&E vs others" classifier to find H&E images in PubMed. Shown replicate AUROC for H&E vs others is 0.9735 for 10-fold (10 replicates of 10-fold has mean±stdev of 0.9746±0.0043) and 0.9549 for LOO (10 reps 0.9547±0.0002), while H&E vs IHC is 0.9967 for 10-fold (10 reps 0.9977±0.0017) and 0.9907 for LOO (10 reps 0.9954±0.0004). For this and other figures, we show the first replicate.

### 4.3.1 Identifying and filtering for H&E images

We ran increasingly difficult tests using increasingly sophisticated machine learning methods. Our first question is the most basic, but arguably the most important: can machine learning distinguish acceptable H&E-stained human pathology images from all others (Figs 4.2A, S4.14, S4.15)? We show acceptable H&E-stained human pathology images can be distinguished from other images – e.g. natural scenes or different histochemistry stains (Fig 4.7 *at left*) with high performance (AUROC 0.95). Because of the high performance of this classifier, it can be used to partially automate one of our manual data curation tasks, e.g.

identifying acceptable images on social media. More importantly, when confronted with over one million PubMed articles, we apply this classifier to filter out all the articles that do not have at least one H&E image. To our knowledge, this is the first H&E image detector to filter PubMed articles. PubMed figures increase our searchable dataset by over an order of magnitude, without any additional manual curation effort. Only with a large dataset may we expect to successfully search for rare diseases, and we currently have 126,787 searchable images. This task also serves as a positive control.

### 4.3.2 Distinguishing common stain types

H&E and IHC stain types are the most common in our dataset and are common in practice. We therefore ask if machine learning can distinguish between these stain types, which vary in coloration (Fig 4.2A). Indeed, the classifier performs very well at this discrimination (AUROC 0.99, Fig 4.7 *at right*). Thus, although IHC coloration can vary between red and brown, machine learning can still successfully differentiate it from H&E. *Intra-stain diversity* explains further (Sec S4.5.3). A well-performing classifier such as this can be useful with large digital slide archives that contain a mixture of H&E and IHC slides that lack explicit labels for staining information. Our classifier can automatically and accurately distinguish these stains, so that downstream pipelines may process each stain type in a distinct manner.

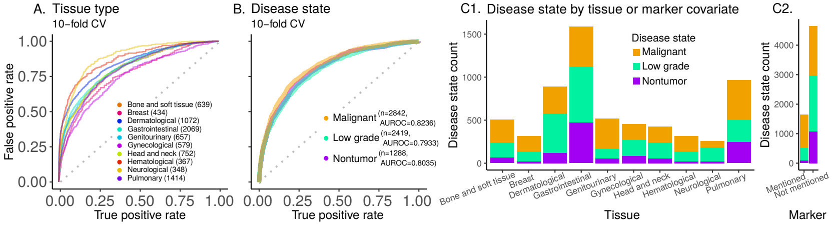
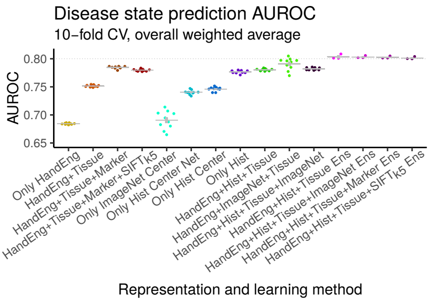


Figure 4.8: **10-tissue type and 3-disease state prediction performance and counts.** **(A)** Classifier performance for predicting histopathology tissue type (10 types, 8331 images). **(B)** Classifier performance for predicting disease state (3 disease states; 6549 images). Overall AUROC is the weighted average of AUROC for each class, weighted by the instance count in the class. Each panel **(A and B)** shows AUROC (with tenfold cross-validation) for the chosen classifier. Random Forest AUROC for tissue type prediction is 0.8133 (AUROC for the ten replicates: mean±stdev of 0.8134±0.0007). AUROC is 0.8085 for an ensemble of our deep-learning-random-forest hybrid classifiers for disease state prediction (AUROC for the ten replicates: mean±stdev of 0.8035±0.0043). **(C1)** Disease state counts per tissue type. The proportion of nontumor vs. low grade vs. malignant disease states varies as a function of tissue type. For example, dermatological tissue images on social media are most often low grade, but malignancy is most common for genitourinary images. **(C2)** Disease state counts as a function of whether a marker test (e.g. IHC, FISH) was mentioned (~25% of cases) or not. IHC is the most common marker discussed and is typically, but not necessarily, used to subtype malignancies.

### 4.3.3 Distinguishing ten histopathology tissue types

We next ask if machine learning can distinguish the ten tissue types present in our Twitter dataset (Fig 4.2B). *Tissue hashtags and keywords* discusses this further (Sec S4.5.6). The tissue types were distinguishable (AUROC 0.81, Fig 4.8A) and, as expected, this task was more difficult than stain-related tasks. Being able to identify tissue types may help to detect contaminating tissue in a slide.



| Method | mean | stdev |
|---|---|---|
| Only HandEng | 0.6843 | 0.0012 |
| HandEng+Tissue | 0.7515 | 0.0018 |
| HandEng+Tissue+Marker | 0.7846 | 0.0023 |
| HandEng+Tissue+Marker+SIFT$_{k5}$ | 0.7796 | 0.0019 |
| Only Hist Center ImageNet | 0.6907 | 0.0163 |
| Only Hist Center Net | 0.7402 | 0.0039 |
| Only Hist Center | 0.7461 | 0.0038 |
| Only Hist | 0.7766 | 0.0027 |
| HandEng+Hist+Tissue | 0.7803 | 0.0019 |
| HandEng+ImageNet+Tissue | 0.7909 | 0.0108 |
| HandEng+Hist+Tissue+ImageNet | 0.7821 | 0.0021 |
| HandEng+Hist+Tissue Ens | **0.8035** | **0.0043** |
| HandEng+Hist+Tissue+ImageNet Ens | **0.8031** | **0.0022** |
| HandEng+Hist+Tissue+Marker Ens | **0.8025** | **0.0021** |
| HandEng+Hist+Tissue+SIFT$_{k5}$ Ens | **0.8014** | **0.0022** |

Figure 4.9: **Disease state prediction performance for machine learning methods.** For deep learning we use a ResNet-50. For shallow learning we use a Random Forest. We train a Random Forest on deep features (and other features), to combine deep and shallow learning (Fig 4.3C *top*). Error bars indicate standard error of the mean. Points indicate replicates. Gray lines indicate means. Performance increases markedly when including tissue type covariate for learning (even though tissue type is missing for some patients), when using deep learning to integrate information throughout entire image rather than only the center crop, and when using an ensemble of classifiers. Performance exceeds AUROC of 0.8 (*at right*). We conclude method xii ("HandEng+Hist+Tissue Ens") is the best we tested for disease state prediction, because no other method performs significantly better and no other simpler method performs similarly. Methods are, from left to right, (i) Random Forest with 2412 hand-engineered features alone for 512×512px scaled and cropped center patch, (ii) Random Forest with tissue covariates, (iii) Random Forest with tissue and marker covariates, (iv) method iii additionally with SIFT$_{k5}$ features for Random Forest, (v) only-natural-image-trained ResNet-50 at same scale as method i with center 224×224px center patch and prediction from a Random Forest trained on 2,048 features from the ResNet-50 (Fig 4.3) (vi) histopathology-trained ResNet-50 at same scale as method i with center 224×224px center patch and prediction from top 3 neurons (Fig 4.3B *top*), (vii) histopathology-trained ResNet-50 with Random Forest trained on 100 features from 224×224px center patch per method vi, (viii) histopathology-trained ResNet-50 features at 21 locations throughout image summed and Random Forest learning on this 100-dimensional set representation with 2,412 hand-engineered features, (ix) method viii with tissue covariates for histopathology-trained ResNet-50 and 2,412 hand-engineered features for Random Forest learning (i.e. Fig 4.3C sans marker information), (x) method ix with an only-natural-image-trained ResNet-50 instead of a histopathology-trained ResNet-50 for Random Forest learning, (xi) method ix with both an only-natural-image-trained ResNet-50 and a histopathology-trained ResNet-50 for Random Forest learning, (xii) method ix with an ensemble of three Random Forest classifiers such that each classifier considers an independent histopathology-trained ResNet-50 feature vector in addition to 2,412 hand-engineered features and tissue covariate, (xiii) method xii where each Random Forest classifier in ensemble additionally considers only-natural-image-trained ResNet-50 features, (xiv) method xii where each Random Forest classifier in ensemble additionally considers the marker mention covariate (i.e. this is an ensemble of three classifiers where Fig 4.3C is one of the three classifiers), (xv) method xii where each Random Forest in ensemble additionally considers SIFT$_{k5}$ features for learning.

### 4.3.4 Deep learning predicts disease state across many tissue types

Pathologists routinely make decisions about whether a tissue shows evidence of nontumoral disease, low grade disease, or malignant disease, while ignoring spurious artifacts (Fig S4.11). We therefore ask whether machine learning can perform well on this clinically important task. For this, we use our most common stain type, H&E, including only those images that are single-panel and deemed acceptable (Fig S4.14). We systematically test increasingly sophisticated machine learning methods (Fig 4.9) with the goal of achieving the highest possible performance. The simplest baseline model we consider, a Random Forest on the 2,412 hand-engineered features (Fig S4.19), achieves an AUROC of 0.6843±0.0012 (mean±stdev, Fig 4.9). Conversely, an ensemble of our deep-learning-random-forest hybrid classifiers achieves much higher performance, with AUROC 0.80 (Fig 4.9). To our knowledge, this is the first classifier that predicts the full spectrum of disease states, i.e. nontumor, low grade, and malignant (Figs 4.2, 4.8B, 4.9).

### 4.3.5 Texture and tissue are important clinico-visual features of disease

We next determine which features are important to our machine learning classifier for disease state prediction. To do this, we interpret the Random Forest feature importance to gain insight into the clinico-visual features that are predictive of disease state. Our analyses suggest that texture (e.g. Local Binary Patterns)

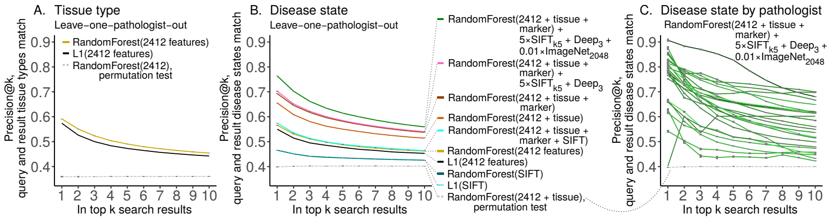


Figure 4.10: **Case similarity search performance.** We report search performance as precision@k for leave-one-pathologist-out cross-validation for **(A)** tissue and **(B)** disease state. We note search based on SIFT features performs better than chance, but worse than all alternatives we tried. Marker mention information improves search slightly, and we suspect cases that mention markers may be more relevant search results if a query case also mentions markers. SIFT$_{k5}$ and histopathology-trained Deep$_3$ features improve performance even less, but only-natural-image-trained ImageNet$_{2048}$ deep features increase performance substantially (Table S4.1). **(C)** We show per-pathologist variability in search, with outliers for both strong and weak performance. Random chance performance is shown as a dashed gray line. In our testing, performance for every pathologist is always above chance, which may suggest performance will be above chance for patient cases from other pathologists. We suspect variability in staining protocol, variability in photography, and variability in per-pathologist shared case diagnosis difficulty may underlie this search performance variability. The pathologist where precision@k=1 is lowest shared five images total for the disease prediction task, and these images are of a rare tissue type. Table S4.2 shows per-pathologist performance statistics.

and color (e.g. Color Histograms) features are most important for pathology predictions and search, followed by the tissue type clinical covariate (Fig 4.4). *Marker mention and SIFT features excluded from Random Forest feature importance analysis* discusses further (Sec S4.5.10). Our method is therefore multimodal, in that it learns from both visual information in the images and their associated clinical covariates (e.g. tissue type and marker mention). Both modalities improve search performance, as discussed in the following section.

### 4.3.6 Disease state search, first pan-tissue pan-disease method

In light of pathology-agnostic approaches to pathology search[88,89], we ask if pathology-specific approaches to pathology search may perform better. Indeed, search is the main purpose of our social media bot. Moreover, others have noted task-agnostic features may suffer from poorly understood biases, e.g. features to distinguish major categories (e.g. cats and dogs) in natural images may systematically fail to distinguish major categories in medical images (e.g. ophthalmology or pathology)[106]. To evaluate performance of search, we show precision@k for k=1…10 (Fig 4.10). As a positive control, we first test search for similar tissues (Fig 4.10A), e.g. if the search query image is breast pathology then the top search results should be breast pathology. Here, precision@k=1 = 0.6 means 60% of the time the search query image and top search result image have matching tissue types, e.g. both are breast, or both are gastrointestinal, etc. We subsequently test search for similar disease states (Fig 4.10B, Table S4.1), e.g. if the search query image is malignant then the top search results should be malignant. Here, precision@k=1 = 0.76 means 76% of the time the search query image and top search result image have matching disease states (e.g. both malignant, both nontumor, etc), while precision@k=8 = 0.57 means the search query image matches 57% of the top 8 search results, i.e. 4-5 of the top 8 search results are malignant when the search query image is malignant. To estimate performance in general for each method, we perform 10 replicates of leave-one-pathologist-out cross-validation with different random seeds (i.e. 0,1,…,9). This allows variance to be estimated for Random Forest learning, but methods based exclusively on the L1 norm are fully deterministic, so these have zero estimated variance (Table S4.1). We follow two-sample hypothesis testing, where one set of 10 replicates is compared to a different set of 10 replicates. To calculate a U statistic and a p-value, we use the two-tailed Wilcoxon rank-sum test on precision@k=1, which tests for significant differences in precision for the first search result on average. For search's statistical null model, we train a Random Forest on images with randomly shuffled class labels and evaluate precision@k, as a permutation test (i.e. "RandomForest(2412 + tissue), permutation test" precision@k=1 = 0.3967±0.0044 in Table S4.1, shown in Fig 4.10B). We conclude search performs significantly better than chance (0.7618±0.0018 vs 0.3967±0.0044, $U = 100$, $p = 0.0001817$) and offer specifics below.

Results for disease state search are detailed in *supplementary disease state search results* (Sec S4.5.13). Here, we briefly describe four main findings. First, *clinical covariates improve search performance* (Sec S4.5.13). Both tissue type (0.5640±0.0024 vs 0.6533±0.0025, $U = 100, p = 0.0001796$) and marker mention (0.6533±0.0025 vs 0.6908±0.0021, $U = 100, p = 0.0001796$) covariates significantly improve search performance. This suggests that for search these clinical features provide disease state information above and beyond the visual characteristics we have of each image. Second, *in the context of other features, nuclear features of disease are better represented by the most prevalent SIFT clusters rather than all SIFT* (Sec S4.5.13), and the effect of SIFT clusters on search is small but significant (0.6908±0.0021 vs 0.6935±0.0029, $U = 19.5$, $p = 0.02308$). This indicates nuclear features, as represented by SIFT, provide limited but complementary disease-related information for search. Third, *deep features synergize with other features, informing search more than nuclear SIFT features, but less than clinical covariates* (Sec S4.5.13). Specifically, deep features improve search performance less than tissue type (0.5720±0.0036 vs 0.6533±0.0025, $U = 0$, $p = 0.0001806$) and less than marker mentions (0.6602±0.0022 vs 0.6908±0.0021, $U = 0$, $p = 0.0001817$), but more than SIFT clusters (0.6983±0.0016 vs 0.6948±0.0032, $U = 83.5$, $p =$

0.01251). Fourth, *deep features trained only on natural images outperform hand-engineered features for search, and offer best performance when combined with other features* (Sec S4.5.13). Particularly, in the context of clinical covariates, ImageNet$_{2048}$ features demonstrate high importance by offering better search performance than the 2,412 hand-engineered features, SIFT$_{k5}$ features, and histopathology-trained Deep$_3$ features *combined* (0.7517±0.0025 vs 0.7006±0.0026, $U = 100$, $p = 0.0001817$) – although this may change as more data become available or more advanced methods are used. Moreover, we found that adding only-natural-image-trained ImageNet$_{2048}$ deep features to our best-performing model (incorporating hand-engineered features, tissue type, marker mention, SIFT$_{k5}$ features, and Deep$_3$ features) improved search performance further (0.7006±0.0026 vs 0.7618±0.0018, $U = 0$, $p = 0.0001817$), and was the best-performing search method we measured. Taken together, we conclude (i) texture and tissue features are important, (ii) histopathology-trained deep features are less important, (iii) nuclear/SIFT features are least important for disease state search, and (iv) in the context of clinical covariates the only-natural-image-trained ImageNet$_{2048}$ deep features are the most important visual features we tested for search.

## 4.4 Discussion

### 4.4.1 Summary

Pathologists worldwide reach to social media for opinions, often sharing rare or unusual cases, but replies may not be immediate, and browsing potentially years of case history to find a similar case can be a time-consuming endeavor.

Therefore, we implemented a social media bot that in real-time searches for similar cases, links to these cases, and notifies pathologists who shared the cases, to encourage discussion. To facilitate disease prediction and search, we maintain a large pathology-focused dataset of 126,787 images with associated text, from pathologists and patients the world over. This is the first pan-tissue, pan-disease dataset in pathology, which we will share with the community through `pathobotology.org` to promote novel insights in computational pathology. After performing stain- and tissue-related baselines with a Random Forest, we performed a number of analyses on this dataset for disease state prediction and search. To accomplish this, we developed a novel synthesis of a deep convolutional neural network for image set representations and a Random Forest learning from these representations (Figs 4.3, S4.24). We found this model can classify disease state with high accuracy, and be repurposed for real-time search of similar disease states on social media. This interpretable model, combined with its social media interface, facilitates diagnoses and decisions about next steps in patient care by connecting pathologists all over the world, searching for similar cases, and generating predictions about disease states in shared images. Our approach also allowed us to make a number of important methodological advances and discoveries. For example, we found that both image texture and tissue are important clinico-visual features of disease state – motivating the inclusion of both of feature types in multimodal methods such as ours. In contrast, we find deep features trained only on natural images (e.g. cats and dogs) substantially improve search performance, while pathology-specific deep features and cell nuclei features improve less, although combining all these performed best. Finally, we provide important technical advances, because our novel deep feature regularization and activation functions yield approximately binary features and set representations that may be applicable to other domains. In sum, these advances readily translate to patient care by taking advantage of cutting-edge machine learning approaches, large and diverse datasets, and interactions with pathologists worldwide.

### 4.4.2 Comparison with prior studies

Our approach builds on, but greatly extends, prior work in the field of computational pathology. We comment on this briefly here, and describe more fully in *supplementary comparison with prior studies* (Sec S4.5.16). First, much of prior work involves a subset of tissue types or disease states[107,108,109]. However, our study encompasses diverse examples of each. Second, prior studies investigating pathology search take a variety of pathology-agnostic approaches, e.g. (i) using neural networks that were not trained with pathology data[88,89] or (ii) using scale-invariant feature transform (SIFT) features[89,110,111] that do not represent texture or color[112]. Our inclusive approach is different, building a search method for pathology data represented by thousands of features – including SIFT clusters, neural networks, other visual features, and clinical covariates. Our model outperforms pathology-agnostic baselines.

Prior work has found texture and/or color to be important for tissue-related tasks in computational pathology[113,114,115]. We find texture and color to be important for disease-related tasks. Additionally, we go a step further by comprehensively considering the relative contributions of many clinico-visual features to the prediction and search of disease. Such important features include texture, color, tissue type, marker mentions, deep features, and SIFT clusters.

### 4.4.3 Caveats and future directions

Below we discuss the primary caveats (also see *supplementary caveats* in Sec S4.5.17) and future directions (also see *supplementary future directions* in Sec S4.5.18).

**Diagnosis disagreement or inaccuracy** First, there is a risk of error in our data because many different pathologists share cases, and they may disagree on the most appropriate hashtags or diagnosis. Moreover, there may be diagnostic inaccuracies from the pathologist who posted the case, or other pathologists. We find these situations to be rare, but if they occur, the case tends to have an increased amount of discussion, so we can identify these situations. Second, our nontumor/low-grade/malignant keyword rules may be incorrect or vague. For these first and second caveats, we take a majority vote approach, manually curate as needed, and discuss. Indeed, as we discussed amongst ourselves the hyperplasia in Fig 4.5, it became clear we needed to explicitly mention preneoplastic disease is included in the low grade disease state category.

**Dataset case sampling and region of interest biases** Our dataset may have both (i) a case sampling bias and (ii) a region of interest sampling bias. First, there may be case sampling bias if we typically have unusual cases that pathologists consider worth sharing, and our cases by necessity only come from pathologists on social media. We plan to advocate sharing of normal tissue and less unusual cases to circumvent this bias. Second, the pathologist who shares the case chooses which images to share, typically sharing images of regions of interest that best illustrate the diagnosis, while ignoring other slides where the diagnosis is less straightforward. In future work, we will include whole slide images for additional context.

**Dataset size and granularity** To increase the granularity and accuracy of tissue type predictions, we first plan to expand the size of this dataset by recruiting more pathologists via social media, aiming to have representative images for each organ. There are many organs within the gastrointestinal tissue type, for instance. Additionally, we expect our dataset to broaden, including more social media networks and public pathology resources such as TCGA, with our bot integrating these data for search and predictions.

## Conclusion

We believe this is the first use of social media data for pathology case search and the first pathology study prospectively tested in full public view on social media. Combining machine learning for search with responsive pathologists worldwide on social media, we expect our project to cultivate a more connected world of physicians and improve patient care worldwide. We invite pathologists and data scientists alike to collaborate with us to help this nascent project grow.

## Acknowledgments

A.J.S. thanks Dr. Marcus Lambert and Pedro Cito Silberman for organizing the Weill Cornell High School Science Immersion Program. A.J.S. thanks Terrie

Wheeler and the Weill Cornell Medicine Samuel J. Wood Library for providing vital space for A.J.S., W.C., and S.J.C. to work early in this project. A.J.S. thanks Dr. Joanna Cyrta of Institut Curie for H&E-saffron (HES) discussion. A.J.S. thanks Dr. Takehiko Fujisawa of Chiba University for his free pathology photos contributed to social media and this project via @Patholwalker on Twitter.

A.J.S. was supported by NIH/NCI grant F31CA214029 and the Tri-Institutional Training Program in Computational Biology and Medicine (via NIH training grant T32GM083937). This research was funded in part through the NIH/NCI Cancer Center Support Grant P30CA008748.

S.Y. is a consultant and advisory board member for Roche, Bayer, Novartis, Pfizer, and Amgen – receiving an honorarium.

T.J.F. is a founder, equity owner, and Chief Scientific Officer of Paige.AI.

We are grateful to the patients who made this study possible.

## Author Contributions

**Conceptualization:** AJS, MA.

**Data curation:** AJS, WCJ, SJC, LGP, BSP, MPP, NZ, BDS, SY, AM.

**Methodology, Software, Validation, Formal analysis, Investigation, Writing (original draft):** AJS.

**Funding acquisition:** AJS, TJF.

**Project administration:** AJS, WCJ, SJC, MA, TJF.

**Resources (pathology) and discussion:** LGP, BSP, MPP, RSS, KH, NZ, BDS, SY, BX, SRA, AM, KAJ, KRO, SM, CM, HY, YR, RHA, OOF, JMG, CR, CS, JG, DES.

**Resources (computational):** AJS, TJF.

**Supervision:** MA, TJF.

**Visualization, wrote annotation files:** AJS, WCJ, SJC.

**Writing (editing):** AJS, MA.

**Writing (reviewing):** AJS, LGP, BSP, MPP, SY, BX, AM, OOF, JMG, CS, SJS, MA, TJF.

**Answered annotator questions:** LGP, BSP, MPP, NZ, BDS, SY, BX, SRA, AM, SM, CM, HY, YR, RHA, OOF, CS, SJS.

## S4.5 Supporting information

### S4.5.1 Image data overview

The goal of obtaining images from practicing pathologists worldwide is to create a dataset with a diverse and realistic distribution of cases. A worldwide distribution (Fig 4.1A) may be appropriate to overcome potential biases inherent at any single institution, such as stain chemistries or protocols. Our dataset includes a wide variety of stains and techniques (Fig 4.2A) – even variety for a single stain, e.g. H&E stains (Fig S4.15). H&E stain composition may vary by country – e.g. in France, H&E typically includes saffron, which stains collagen fibers. Phyloxin may be used instead of eosin. This helps differentiate between connective tissue and muscle, or to see cell cytoplasm better against a fibrous background. This stain may be referred to as "HES" or "HPS", and we consider it H&E. *Intra-stain diversity* discusses further (Sec S4.5.3). Our dataset includes gross sections (Fig S4.13) that pathologists share alongside images of stained slides. In addition to variation in the signal of interest (i.e., stain, tissue, or disease), we find variability in the noise (i.e. pathology artifacts, Fig S4.11). Such noise may initially seem undesirable, but is likely important for machine learning techniques to robustly predict which image motifs are relatively unimportant rather than prognostic. Finally, our dataset includes a variety of parasites and other [micro]organisms (Fig S4.12, and Fig S4.15A,E), an important consideration in developing countries.

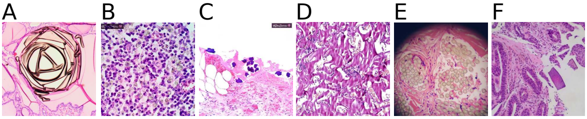


Figure S4.11: **Artifact and foreign body images.** Our dataset includes artifacts and foreign bodies which machine learning should not consider prognostic. All panels human H&E. **(A)** *B.X.*: Colloid. **(B)** *L.G.P.*: Barium. **(C)** *L.G.P.*: Oxidized regenerated cellulose, a.k.a. gauze, granuloma may mimic mass lesion[116]. **(D)** *R.S.S.*: Hemostatic gelatin sponge, a.k.a. Spongostan$^{TM}$, may mimic necrosis. **(E)** *S.Y.*: Sutures, may mimic granuloma or adipocytes. **(F)** *L.G.P.*: Crystallized kayexelate, may mimic mass lesion or parasite.

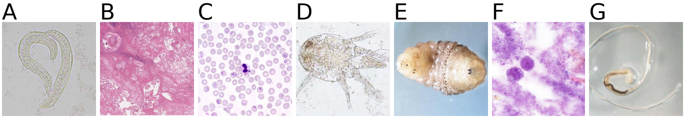


Figure S4.12: **Parasitology images.** Our dataset includes diverse parasitology samples. **(A)** *B.S.P.*: *Strongyloides stercoralis*, light microscopy. **(B)** *B.S.P.*: *Dirofilaria immitis*, in human, H&E stain. **(C)** *B.S.P.*: *Plasmodium falciparum*, in human, Giemsa stain. **(D)** *B.S.P.*: Incidental finding of unspecified mite in human stool, light microscopy. **(E)** *B.S.P.*: *Dermatobia hominis*, live gross specimen. **(F)** *B.S.P.*: *Acanthamoeba*, in human, H&E of corrective contact lenses. **(G)** *B.S.P.*: *Trichuris trichiura*, gross specimen.

### Defining an acceptable pathology image

To create our pathology social media database, we first identified pathology images, and second, narrowed down the set of pathology images into those that were of sufficient quality to be used and could be shared publicly. By "pathology image", we mean images that a pathologist may see in clinical practice, e.g. gross sections, microscopy images, endoscopy images, or X-rays. An image designated as a "pathology image" is not necessarily an image of diseased tissue. After we identified pathology images, we screened them for inclusion in our dataset. "Acceptable images" are those that do not meet rejection or discard criteria defined below. If an acceptable image is personally identifiable or private

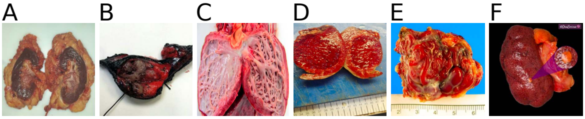


Figure S4.13: **Gross images.** Gross sections are represented in our dataset, putting the slide images in context. **(A)** *M.P.P*: Urothelial carcinoma. **(B)** *M.P.P.*: Lung adenocarcinoma. **(C)** *S.R.A.*: Barth syndrome. **(D)** *N.Z.*: Enlarged spleen. **(E)** *S.R.A.*: Arteriovenous malformation. **(F)** *L.G.P.*: Kidney adrenal heterotopia.

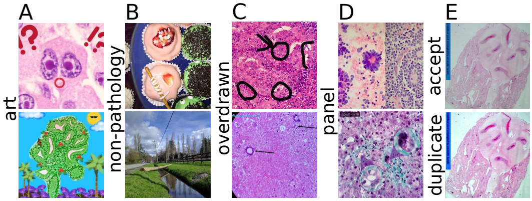


Figure S4.14: **Image acceptability criteria.** Examples of images that are rejected, because they are not pathology images that a pathologist would see in clinical practice. **(A)** *top M.P.P., bottom B.D.S*: “art” rejects. **(B)** *top B.S.P., bottom S.Y.*: “non-pathology” rejects. **(C)** *top B.X., bottom A.M.*: “overdrawn” rejects. **(D)** *top S.R.A., bottom L.G.P.*: “panel” is rejected for some tasks, e.g. H&E vs IHC or disease state prediction, but not for others, e.g. H&E vs others. The H&E vs others task retains multi-panel images because multi-panel images that include an H&E panel should be included in our PubMed search results, and this classifier is used to filter PubMed. **(E)** *top and bottom S.R.A.*: top is acceptable H&E (see Sec S4.5.1 for definition), bottom is “dup” [duplicate] rejection.

(see criteria below), we retain the image for some machine learning analyses, but do not distribute the image publicly [for legal reasons].

**Criteria for rejected, discarded, private, or acceptable images** For our manual data curation process, we defined several rejection criteria (Fig S4.14), detailed in Section S4.5.2. Figure S4.14A shows examples of images rejected as “art”, because they are artistically manipulated H&E pathology microscopy images. Figure S4.14B shows examples of images rejected as “non-pathology”,

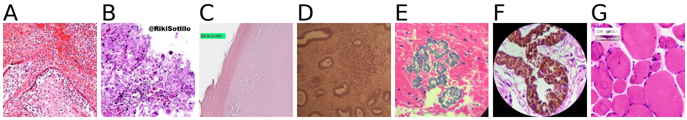


Figure S4.15: **H&E images.** Our dataset includes diverse H&E-stained slide microscopy images. **(A)** *S.R.A.*: Acute villitis due to septic *Escherichia coli*. **(B)** *R.S.S.*: Garlic. **(C)** *R.S.S.*: "Accellular" leiomyoma after ulipristal acetate treatment. **(D)** *R.S.S.*: Brownish appearance from dark lighting. **(E)** *K.R.O.*: *Sarcina* in duodenum. **(F)** *B.D.S.*: Mature teratoma of ovary, pigmented epithelium. **(G)** *K.A.J.*: Central core myopathy.

e.g. parasitology-inspired cupcakes (*top*) and a natural scene image (*bottom*). Non-pathology images are relatively common on pathologists' social media accounts, though we try to minimize their frequency by recruiting pathologists who primarily use their accounts for sharing and discussing pathology. Figure S4.14C shows examples of images rejected as "overdrawn". Overdrawn images are those that have hand-drawn marks from a pathologist (which pathologists refer to as "annotations"), which prevent us from placing a sufficiently large bounding box around regions of interest while still excluding the hand-drawn marks. Section S4.5.2 discusses our "overdrawn" criterion in detail. Figure S4.14D shows examples of images rejected as "panel", because they consist of small panels (*top*) or have small insets (*bottom*); splitting multi-panel images into their constituent single-panel images would substantially increase our manual curation burden. Figure S4.14E *top* is an acceptable H&E-stained pathology image. Figure S4.14E *bottom* is rejected as a duplicate of the S4.14E *top* image, though the colors have been slightly modified, and the original image is a different size.

### S4.5.2 Supplementary Image processing

#### Criteria details for rejected, discarded, private, or acceptable images

Though criteria are outlined in *criteria for rejected, discarded, or acceptable images* (Sec S4.5.1) – more formally, we reject the following image types, during our manual data curation process:

1. Non-pathology images, such as pictures of vacations or food.
2. Multi-panel images, such as a set of 4 images in a 2×2 grid. Images with insets are also rejected. For "H&E vs IHC", "tissue type", and "disease state" tasks, we only accept single-panel images, and leave for future work the complexities of splitting multi-panel images into sets of single-panel images. We accept multi-panel images for the "H&E vs others" task, because we use the classifier trained for this task to filter PubMed, and many H&E images in PubMed are multi-panel, which are useful as search results. Multi-panel images may have black dividers, white dividers, no dividers, square insets in a corner, or floating circular insets somewhere in the image. There may be two or more panels/insets. Per-pixel labels for each panel may be the best solution here, and would support a machine learning approach to split multi-panel images to reduce this additional manual data curation burden.
3. Overdrawn images, where a 256×256px region could not bound all regions of interest in an image. This occurs most frequently if a pathologist draws by hand a tight circle around a region of interest, preventing image analysis on the region of interest in a way that completely avoids the hand-drawn marks.

4. Images that manipulate pathology slides into artistic motifs, such as smiley faces or trees. In contrast, a picture of a painting would be a non-pathology image.

Moreover, we completely discard from analysis certain types of images:

1. Duplicate images, according to identical SHA1 checksums or by a preponderance of similar pixels. However, duplicate images may be shown in search results, if the images are contained in different tweets, because there may be different replies to these tweets as well.
2. Corrupt images, which either could not be completely downloaded or employed unusual JPEG compression schemes that Java's ImageIO[3] library could not open for reading.
3. Pathology images that are owned by pathologists who have not given us explicit written permission to use their images. Consider the following example. When a pathlogist gives us permission to download data, our software bot downloads thousands of that pathologists's social media posts regardless if some of the images in those posts are actually owned by a different pathologist who did not give us permission. We detect these cases when we manually curate the pathologist's data, and discard these images belonging to pathologists who have not given us permission. To elaborate, pathology images that are taken by pathologists and shared on social media are treated the same way as pathology images taken from case reports or copyrighted manuscripts, i.e. if the pathologist or publisher has not provided us explicit written permission to use the image, we discard the pathology image and do not use it.

---

[3] ImageIO documentation available here: `https://docs.oracle.com/javase/7/docs/api/javax/imageio/ImageIO.html`

Images that are not rejected or discarded are deemed “acceptable” pathology images. However, for legal reasons, we cannot distribute all of the images we have from social media, namely:

1. Pathology images obtained from children (including fetuses), which may be identifiable. The data shared on social media are anonymized; thus, we do not have contact information for the child’s parent and therefore cannot obtain consent to distribute a picture of e.g. a child’s X-rays or gross specimens. Although unlikely to be identified by the parent if these images were made public, we prefer to err on the side of caution. However, microscopy slide images are not personally identifiable, so we may distribute these.
2. Personally identifiable pictures involving adults, because they have the right to consent or not to their likeness being distributed. We consider faces, body profiles, automobile license plates, etc to all be personally identifiable pictures involving adults, especially because these data may be cross-referenced against timestamp, location, clinician, institution, medical condition, other people in the picture, etc.
3. Copyrighted content, which includes images of copyrighted manuscripts, pictures of slideshow presentations, and pictures of any brand or logo. A lab picture that includes boxes bearing logos would be a non-pathology image that we cannot distribute, because we do not have permission to distribute any images with the protected logos. A picture of a powerpoint slide at a conference that shows some text outlining a new way to make a clinical decision would also be a non-pathology image that we hold privately and do not distribute. We similarly hold privately an image of text taken from a non-open-access manuscript because it may not be possible

to identify the original source to provide a proper citation, and even if we could, this poses an additional data curation burden that we would rather avoid. Moreover, we prefer to err on the side of caution and not distribute these images, rather than rely on "fair use" or similar law that may expose us to legal challenges and costs[4]. By retaining these images privately, we can train a machine learning classifier to detect these types of images and potentially reduce our manual data curation burden.

### Overdrawn rejection criterion

Here we discuss the details of rejecting images as "overdrawn". Figure S4.14C *top* is rejected as "overdrawn", because the regions of interest (ROIs) in the H&E image that the pathologist refers to in the social media post's text have hand-drawn circles and arrows such that it is not possible to place a 256×256px square over all ROIs without including these circle and arrow marks. We chose 256×256px because deep convolutional neural networks in computational pathology[11] typically require 227×227px (i.e. AlexNet[50] or CaffeNet[49]) or 224×224px (i.e. ResNet[62]) images, and we have used these sizes in the past[67,5]. We note the Inception[117] family of deep convolutional neural networks takes a 299×299px image input, which is larger than 256×256px and is also frequently used in computational pathology[11]. Ideally, each image would have ROIs and hand-drawn arrows/circles annotated at the pixel level, so each image could be annotated as "overdrawn" to arbitrary bounding box sizes, whether 256×256px or 299×299px, and we leave this to future work. Smaller "overdrawn" bounding boxes may allow more images to pass as acceptable, rather than be rejected. A 256×256px image size allows minor rotations and crops for deep learning data augmentation using 224×224px input image sizes. Minor upsampling and/or image reflection at the image's outer boundaries may allow a 256×256px image to work for 299×299px input image sizes. Figure S4.14C *bottom* is rejected as "overdrawn", because this image was originally 783×720px and the arrow marks prevent us from capturing each of the two indicated regions of interest in their own 256×256px square.

---

[4] Courts in the United States have ruled that images posted to social media are still owned by their authors and are not public domain. Indeed, in *Morel v. AFP*, AFP was ordered to pay Morel $1.2 million for copyright infringement because AFP used images that Morel posted to social media.

### Uniform cropping and scaling of original images

Images shared on social media may be any rectangular shape. However, machine learning methods typically require all images be the same size. To accommodate this, we use the following procedure:

1. Take the minimum of two numbers: the original image's height and width.
2. Crop from the center of the original image a square with a side whose length is the minimum from the prior step.
3. Scale this square to 512×512px.

This square is intended to be large enough to represent small details, such as arrows and circles drawn one pixel wide by the pathologist. Such arrows and circles may then be used to predict if an image is "overdrawn" or not. Ideally, the tweet's text would be available alongside the image to give the machine learning the fullest information possible about potential ROIs in the image, for "overdrawn" prediction, but for simplicity here we perform only image-based machine learning.

The motivation for the 256×256px "overdrawn" criterion detailed in Sec S4.5.2 is that there may be an attention layer that scans the original image for 256×256px squares that have no marks from the pathologist. Such marks include circles or arrows for ROI indication, or the pathologist's name to indicate copyright/ownership. Such mark-free 256×256px images may then be used for machine learning on only patient pathology pixels.

### S4.5.3 Data diversity discussion

**Intra-stain diversity**

There is an art and variability in histochemical stains that we have not discussed in the main text, but for completeness mention here. We note that in clinical practice we have observed high variability stains, for instance H&E stains that appear almost neon pink, to GMS stains (discussed below) that had silver (black) deposition throughout the slide. One reason for this is that there are a number of reagents that may be used for staining, each with different qualities that can make the stain darker, brighter, pinker, bluer, etc.

IHC stains typically use an antibody conjugated to a brown stain, namely 3,3′-Diaminobenzidine (DAB). The blue counterstain is typically hematoxylin. However, some laboratories conjugate the antibody to a red stain instead. A small minority of our IHC images are this red variant, which should not be confused with H&E.

There is counterstain variability in Grocotts modification of the Gömöri methenamine silver stain [GMS stain]. Typically the counterstain is green, but a pink counterstain is also available. We may see the pink variant as we acquire more data. Currently we see only green.

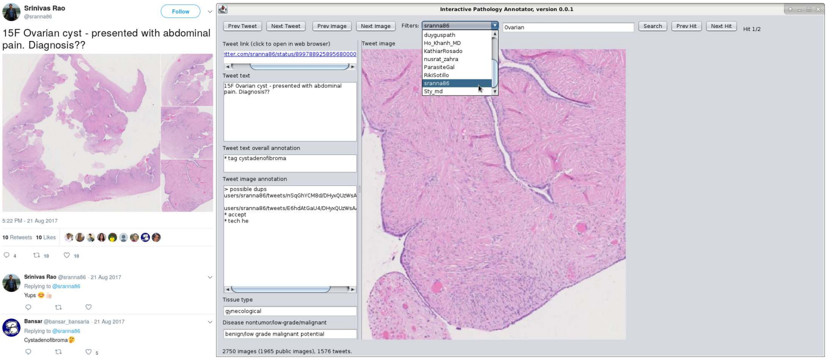
#### Intra-tissue-type diversity

The tissue type hashtags we use are very broad, e.g. #gipath encompasses several organs, such as stomach, small intestine, large intestine, liver, gallbladder, and pancreas. We note, for instance, liver morphology looks nothing like the stomach. Moreover, gynecological pathology, i.e. #gynpath, includes vulva (which looks just like skin, i.e. dermatological pathology, #dermpath), vagina, cervix, uterus, fallopian tubes and ovaries. Again, vulva looks nothing like uterus. A number of tissue features also overlap, such as adipocytes in breast tissue and adipocytes in the subcutaneous fat layer in skin. The amount and distribution of adipocytes typically differs between these tissues however. However, a lipoma in any tissue has a great deal of adipocytes and should not strictly be confused with breast tissue. For all these motivating reasons, we have a future direction to sample every organ within a tissue type hashtag category, for all tissue type hashtag categories.

### S4.5.4 Procedure overview

#### Consent, data acquisition, curation, and review

We follow the procedure outlined in Fig 4.1, and we first obtain data in steps A-D. In step A, we find pathologists on social media (Twitter) who share many pathology cases, or share infrequently shares tissues, e.g. neuropathology. In step B, we contact the pathologist via social media and ask for permission to use their cases. In step C, we download the consenting pathologist's cases shared on social media. In step D, we manually annotate these posted cases for acceptability, e.g. if overdrawn, corrupt, duplicate, multi-panel, art, or non-pathology rejecting per Fig S4.14. *Defining an acceptable pathology image* is explained in further sections. We additionally annotate technique (Fig 4.2A), species (Figs S4.12, S4.15A,B,E), and private status (e.g. personally identifiable pictures of adults or pictures of children). *Image data overview* (Sec S4.5.1) and *criteria for rejected, discarded, private, or acceptable images* (Sec S4.5.1) explain further, e.g. our definition of "overdrawn" or what is [not] pathology. Moreover, if the nontumor/low-grade/malignant status in a tweet is not clear, we read the Twitter discussion thread for this case and manually annotate the case appropriately if possible. Step D also involves clarifying cases that we have trouble annotating, e.g. if it is not clear what stain was used for the image. We first ask the pathologist who posted this case to social media. If we do not obtain an answer from that pathologist, we (i) ask a pathologist at our local institution (i.e. author *S.J.S.*) for an opinion or (ii) ask pathologist coauthors would have shared cases with us. Pathologist validation of tissue and disease labels is an important part of step D, and we use two tools for this. The first tool is the Interactive Pathology Annotator (IPA) (Figs S4.16, S4.17), which pathologists may run on their desktop to browse their case annotations. The second tool is our social media bot "pathobot" (Fig 4.1), which interacts with collaborating pathologists, then publicly posts results of search and disease state predictions. Pathobot search results may indicate annotation issues, e.g. if a bone and soft tissue pathology search returns a breast pathology result, we may check if the breast result was mistakenly labeled as bone and soft tissue. Though we sometimes manually



Srinivas Rao @sranna86

15F Ovarian cyst - presented with abdominal pain. Diagnosis??

5:22 PM - 21 Aug 2017

10 Retweets 10 Likes

Srinivas Rao @sranna86 · 21 Aug 2017
Replying to @sranna86
Yups 😊👍🏻

Bansar @bansar_bansaria · 21 Aug 2017
Replying to @sranna86
Cystadenofibroma🤔

Interactive Pathology Annotator, version 0.0.1

Filters: sranna86

Ovarian

Hit 1/2

Tweet link (click to open in web browser): itter.com/sranna86/status/8997889258956800000

Tweet text: 15F Ovarian cyst - presented with abdominal pain. Diagnosis??

Tweet text overall annotation: * tag cystadenofibroma

Tweet image annotation:
> possible dups
users/sranna86/tweets/nSqGhYCM8d/DHyxQUzWsA

users/sranna86/tweets/E6hdAtGaU4/DHyxQUzWsA
* accept
* tech he

Tissue type: gynecological

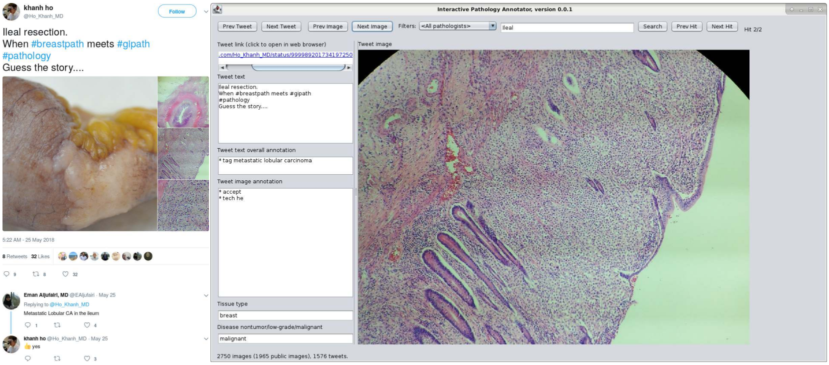
Disease nontumor/low-grade/malignant: benign/low grade malignant potential

2750 images (1965 public images), 1576 tweets.

Figure S4.16: **Interactive Pathology Annotator tool and social media dialogue.** *At left*: Pathologist (author *S.R.A.*) discusses a case. Without mentioning the diagnosis himself, he confirms the diagnosis suggested by a second pathologist, i.e. cystadenofibroma, which we explicitly annotate. *At right*: Our Interactive Pathology Annotator (IPA) tool displays an image from this case, in the context of the tweet overall. IPA is a portal for pathologists to (i) browse tweets and images in the dataset; (ii) validate our data annotations; (iii) check our tissue type categorization algorithm results, (iv) check our nontumor, low grade, and malignant categorization algorithm results; (v) search tweets for specific keywords or diagnoses; (vi) filter out all cases except those from a specific pathologist; and (vii) click the link to the original tweet on Twitter for context.

annotate some cases, most cases are annotated in a crowd sourced fashion. We use social media post hashtags and a pathologist-reviewed rule-based text processing algorithm to determine tissue type and disease state (Fig S4.18).

**Interactive Pathology Annotator discussion** For completeness, we show another example of the use of our Interactive Pathology Annotator (IPA) tool (Fig S4.17), which some pathologists have used to check that tissue and disease state annotations were correct. This is a case of metastatic disease, from breast to gastrointestinal tissue, showing a diffuse pattern of lobular carcinoma that is more common in breast.

 
khanh ho @Ho_Khanh_MD

Ileal resection.
When #breastpath meets #gipath
#pathology
Guess the story....

5:22 AM - 25 May 2018

8 Retweets 32 Likes

Eman Aljufairi, MD @EAljufairi · May 25
Replying to @Ho_Khanh_MD
Metastatic Lobular CA in the ileum

khanh ho @Ho_Khanh_MD · May 25
👍 yes

Interactive Pathology Annotator, version 0.0.1

Filters: <All pathologists>

Ileal

Hit 2/2

Tweet link (click to open in web browser): .com/Ho_Khanh_MD/status/999989201734197250

Tweet text:
Ileal resection.
When #breastpath meets #gipath
#pathology
Guess the story....

Tweet text overall annotation:
* tag metastatic lobular carcinoma

Tweet image annotation:
* accept
* tech he

Tissue type: breast

Disease nontumor/low-grade/malignant: malignant

2750 images (1965 public images), 1576 tweets.

Figure S4.17: **Interactive Pathology Annotator tool and tissue type hashtags.** *At left*: Pathologist (author *K.H.*) discusses case. Without mentioning the diagnosis himself, he confirms diagnoses suggested by other pathologists, i.e. lobular breast carcinoma metastasized to ileum, which we explicitly annotate. *At right*: IPA shows that our tissue type categorization algorithm categorizes this tweet as breast pathology rather than gastrointestinal. The primary tumor is in breast. We define the tissue classification task this way to have applications for tumor site of origin prediction.

### Machine learning, search, checks, and social media bot

Our procedure (Fig 4.1) continues with data analysis in steps E-G. In step E, we use machine learning to train a classifier for a supervised learning task. For example, a task may be to predict the disease state evident in a H&E image: malignant, benign / low grade malignant potential [low grade], or nontumoral pathology. This is a three-class classification task. Our baseline classifier is a Random Forest,[70] which we compare to deep learning. We reuse the classifier to compute a similarity metric for search. In step F, a pathologist posts to social media three key pieces of information together: (i) pathology images, (ii) text descriptions, and (iii) the text "@pathobot". Our social media bot is triggered when mentioned this way, with parts (i) and (ii) forming the pathology search query. In step G, our bot first searches its social media database of cases, then searches its larger PubMed database. A search result will be ranked highly when the query and result tissue types match, when more clinical keywords are shared between the query and result, and when the images are similar. Additionally, the bot will use an ensemble of classifiers to compute with uncertainties the probability of each disease state in each image. This prediction is a sanity check for search, i.e. if the prediction is uncertain or inaccurate, then the search results may be suspect.

### Social media interactions, search, notifications, recruitment

One cycle of our procedure (Fig 4.1) culminates with concluding social media interactions in steps H and I, before ultimately repeating at step A. The social media bot posts its social media search results, PubMed search results, and disease state prediction results. The social media search results include links to similar cases posted to social media. The social media platform may notify pathologists that their posted case has been linked. These pathologists may discuss the putatively similar case. Our bot leverages text information from the pathologist's search post and reply posts. In this way our bot's search is informed by any diagnosis in the differential from any replying pathologist. When multiple pathologists mention the same clinical or diagnostic keywords, those keywords are weighted more highly for search. In effect, search is a collective endeavor by all pathologists in the community discussing the case. The same search repeated over time may be more informed when more pathologist discussion has accumulated over that time. We find that integrating our bot into social media discussions sometimes inspires pathologists to contact us, share with us, and collaborate with us. We then return to step A, and we collect more data for search and classifier training.

### S4.5.5 Text data overview

For supervised learning, we use regular expressions to detect keywords in a tweet's text, to determine labels for the tweet's images. The text and included hashtags may indicate tissue type or disease state.

#### Tissue type categories from text

Prior work has discussed pathology-related hashtags as a way to make pathology more accessible on social media[118]. Pathologists use hashtags to indicate histopathology tissue types, such as "#gynpath" to indicate gynecological pathology (Fig 4.2B). Sometimes alternative spellings are used, such as "#ginpath". Abbreviations are also common, e.g. "#breastpath" and "#brstpath" all mean the same thing: breast pathology (Fig S4.17). A pathology hashtag ontology is available at `https://www.symplur.com/healthcare-hashtags/ontology/pathology/`. Because a tweet can have more than one hashtag, we took the first tissue type hashtag to be the "primary" tissue type of the tweet, and ignored the others. Section S4.5.6 discusses a special case. As detailed in Section S4.5.6, we used hashtags and keywords for all tweets in a message thread to identify the ten tissue types on Twitter, finding 233 bone and soft tissue tweets, 155 breast tweets, 415 dermatological tweets, 794 gastrointestinal tweets, 239 genitourinary tweets, 218 gynecological tweets, 308 head and neck tweets, 115 hematological tweets, and 559 pulmonary tweets.

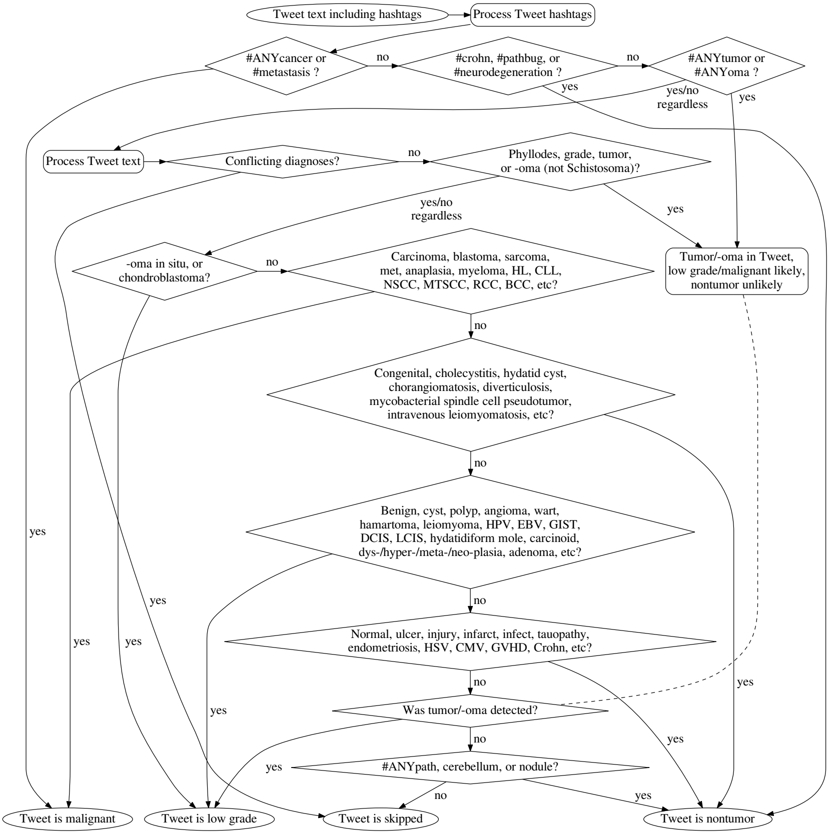

Tweet text including hashtags
Process Tweet hashtags
#ANYcancer or #metastasis ?
no
#crohn, #pathbug, or #neurodegeneration ?
no
#ANYtumor or #ANYoma ?
yes
yes/no regardless
yes
Process Tweet text
Conflicting diagnoses?
no
Phyllodes, grade, tumor, or -oma (not Schistosoma)?
yes/no regardless
yes
-oma in situ, or chondroblastoma?
no
Carcinoma, blastoma, sarcoma, met, anaplasia, myeloma, HL, CLL, NSCC, MTSCC, RCC, BCC, etc?
Tumor/-oma in Tweet, low grade/malignant likely, nontumor unlikely
no
Congenital, cholecystitis, hydatid cyst, chorangiomatosis, diverticulosis, mycobacterial spindle cell pseudotumor, intravenous leiomyomatosis, etc?
no
Benign, cyst, polyp, angioma, wart, hamartoma, leiomyoma, HPV, EBV, GIST, DCIS, LCIS, hydatidiform mole, carcinoid, dys-/hyper-/meta-/neo-plasia, adenoma, etc?
yes
yes
no
Normal, ulcer, injury, infarct, infect, tauopathy, endometriosis, HSV, CMV, GVHD, Crohn, etc?
yes
yes
yes
no
Was tumor/-oma detected?
yes
yes
no
yes
#ANYpath, cerebellum, or nodule?
no
yes
Tweet is malignant
Tweet is low grade
Tweet is skipped
Tweet is nontumor

Figure S4.18: **3-disease text processing algorithm flowchart.** Flowchart of algorithm that processes a single tweet's text to categorize it as nontumor (309 images), benign/low grade malignant potential [low grade] (347 images), or malignant (385 images). A tweet may be skipped (132 images, i.e. 11.3% of images) when the pathologist discusses multiple possible diagnoses for this case or when no pathology keywords are found. Dashed line indicates early steps where tumor/-oma detected, and a later step where detected tumor/-oma considered for possible low grade categorization. Nontumor, low grade, and malignant are defined in Sec S4.5.7. Flowchart steps are detailed in Sec S4.5.7. The algorithm has many steps in order to parse overlapping words that have different diagnoses. For instance, if "Lobular carcinoma in situ of the breast" (which is a low grade disease) was the tweet text, the algorithm has an early step to categorize "carcinoma in situ" as low grade (which is correct here) because a later step categorizes "carcinoma" as malignant (which is not correct here). Indeed, tweet text "Carcinoma of the breast" describes a malignant disease and the algorithm categorizes it malignant because "in situ" is absent. Besides "carcinoma in situ" (low grade) and "carcinoma" (malignant), the algorithm distinguishes "chorangiomatosis" (nontumor) from "angioma" (low grade), "hydatidiform mole" (low grade) from "hydatid cyst" (nontumor) from "ovarian cyst" (low grade) from "cholecystitis" (nontumor), and "intravenous leiomyomatosis" (nontumor) from "leiomyoma" (low grade).

#### Nontumor, low grade, and malignant categories from text

We define three broad disease state categories (Figs 4.2C, S4.18) to use as labels for supervised learning. Our “nontumor” category of 589 tweets includes normal tissue, artifacts, injuries, and nontumoral diseases, e.g. Crohn’s disease, herpes simplex infection, and myocardial infarction. Our “malignant” category of 1079 tweets includes all malignant disease, including carcinoma, blastoma, sarcoma, lymphoma, and metastases. Our definition of malignancy in epithelial cancers is the ability to breach the basement membrane, i.e. a malignant tumor escapes containment and is therefore no longer treatable with surgical resection. Our “pre-neoplastic/benign/low grade malignant potential” [low grade] category of 919 tweets is then all tumors or pre-cancer/neoplastic lesions that are not yet invasive/malignant, e.g. hamartomas, carcinoid tumors, adenomas, and carcinoma *in situ*. Details in Section S4.5.7. For the nontumoral vs low grade vs malignant task, text processing was more complicated than the tissue type task (Sec S4.5.6) because (i) of a heavy reliance on diagnosis keyword matching (flowchart in Fig S4.18), and (ii) additional per-tweet and per-image annotations to clarify nontumor/low-grade/malignant state, which may involve feedback from a pathologist. Details in Section S4.5.7.

### S4.5.6 Supplementary Text processing

#### Hashtag special case

A hashtag special case is “#bstpath”, bone and soft tissue pathology, which we include in our breast pathology category only when the social media post’s text also includes the word “breast” or other breast-related keywords. Such keywords are listed further below in this subsection. Examples of such tweets are *“Pleomorphic lobular carcinoma of the breast: Beautiful cells but nasty tumour #pathology #pathologists #BSTPath”* and *“Now at my desk, W(47y-o) breast nodule...Could be it siliconoma?? But it isn't noted giant cells #pathology #pathologists #BSTpath”*.

### Tissue hashtags and keywords

We found a large number of pathology-related hashtags. We grouped alternative spellings, e.g. #ginpath is #gynpath, #brstpath is #breastpath, and #headandneckpath is #entpath. We also grouped less common hashtags with more common hashtags, e.g. #cardiacpath is #bstpath (bone and soft tissue). Some groupings were broad, e.g. #headandneckpath, #thyroid, #salivary, #oralpath, #endocrine, #endopath, #oralpath, #eyepath, and #ocularpath are all #entpath.

To expand the per-tissue tweet counts, we moved beyond the hashtags and next searched for keywords in the tweet using Perl regular expressions. Further, if a tweet's tissue type could not be determined by hashtags and keywords, we assigned the tissue type of any other tweet in the message thread of tweets. For example, if a tweet of unknown tissue type were a reply to a tweet of known genitourinary type, then we considered both tweets to be genitourinary.

### S4.5.7 Nontumoral, Low grade, and Malignant task details

Tasks involving distinguishing nontumoral disease, low grade tumors, and malignant tumors (Fig 4.2C) are our most difficult tasks. The acknowledged definition of “malignant” in epithelial cancers is the ability to breach the basement membrane, i.e. a malignant tumor escapes containment and is therefore no longer “treatable with surgical resection”. A malignant tumor can invade into the adjacent tissue, lymphatics, and blood vessels. For machine learning, we define a three categories of disease: (a) normal tissue and nontumoral disease; (b) benign, low grade, and oncovirus-driven tumoral disease; and (c) malignant tumors – but there are number of caveats with this, because:

1. there is a spectrum of pathology rather than an oversimplified 3-class nontumoral/low-grade/malignant system.
2. the benign/malignant dichotomy may be more vague in certain tissues e.g. central nervous system (CNS) primary tumors such as chordomas.
3. vague terms like adenoma are typically benign but may be malignant, and likewise vague terms like anaplasia are more often associated with malignancy but not always.
4. vague terms like anaplasia and neoplasia make no real reference to the malignancy of lesions i.e. there are benign anaplastic lesions, while neoplasia is almost synonymous with tumor.
5. terms like tumor do not provide information about benign or malignant state, though normal/nontumoral can be ruled out.
6. there may be some disagreement if some terms, e.g. “carcinoma *in situ*”, are more appropriate to include as low grade, or if instead should be con-

sidered malignant due to their malignant potential or treatment implications. For instance, ductal carcinoma *in situ* (DCIS) typically needs to be removed with surgery or radiotherapy, whereas lobular carcinoma *in situ* (LCIS) typically does not. DCIS's lower grade counterpart, atypical ductal hyperplasia, may get surgery or not. We believe treatment implications are a separate task. Typically, tweets do not include a decision to perform surgery or not, so additional annotations may be needed for the surgery task. We assign all pre-cancer and tumoral disease with malignant potential to the "low grade" category, in light of these benign/malignant ambiguities and data limitations.

7. the diagnosis should be known before deciding benign/malignant, but it is very difficult to know the full diagnosis from the brief, generic, descriptive terms in the tweet.

**Text processing for Nontumoral, Low grade, and Malignant tasks**

To determine if an acceptable H&E human microscopy image is nontumoral, low grade, or malignant, we use regular expressions (Fig S4.18) as we did for tissue type classification. However, keywords differed and we considered all tweets in a message thread per Sec S4.5.5. To infer these message threads of tweets, we downloaded from Twitter each tweet's metadata (in JavaScript Object Notation (JSON) format), which describes the parent tweet for each tweet. If tweet A is a reply to tweet B, then tweet A is the parent of tweet B, and both tweets are in the same message thread.

Our heirarchical algorithm for nontumor/low-grade/malignant keyword-matching shown in Fig S4.18, and details for each step follow. First, to determine if a single tweet indicated nontumoral, low grade, or malignant, we looked for specific hashtags in a tweet's text that indicated malignancy, tumoral status, or nontumoral status. For illustartion, what follows is a subset of our rules, with commentary.

1. Malignant: `/#[a-z]*cancer/i` or `/#metastas[ei]s/i`
    - The first regular expression in this set matches #ANYcancer, where ANY can be any non-whitespace characters, e.g. "#bladdercancer" and "#breastcancer" both match, as well as "#cancer".
    - Metastasis is a sign of malignant cancer, so tweets with #metastasis or #metastases hashtags are malignant.
    - If any matching keyword is detected, no further keyword processing is performed. The tweet is malignant.
2. Nontumoral: `/#crohn/i` or `/#neurodegeneration/i` or `/#pathbug/i`
    - Crohn's disease and neurodegeneration are not tumoral diseases, so this tweet is in the nontumoral/normal category. This `/#crohn/i` regular expression is case-insensitive, so it matches "#crohn", "#Crohn", and "#CROHN". The #pathbug hashtag indicates a parasite or other microorganism is in the image, which is also nontumoral.
    - If any matching keyword is detected, no further keyword processing is performed. The tweet is nontumoral.
3. Tumoral status (ambiguously low grade or malignant): `/#[a-z]*tumou?r/i` or `/#[a-z]*oma/i`
    - The first regular expression in this set matches #ANYtumor or #ANYtumour, where ANY can be any non-whitespace characters, e.g.

“#BrainTumor” and “#phyllodestumour” both match, as well as “#tumor”.

- The second regular expression matches #ANYoma, e.g. #Lymphoma and #leiomyoma both match.
- Because “tumor” and “-oma” do not necessarily mean a tumor is low grade or malignant, further keyword matching is performed. It is unlikely that the tweet is nontumoral. If no other specific information is found after all further keyword matching is performed, the tumor is presumed to be low grade.

Second, if no hashtags matched, we then analyzed keywords in the tweet text.

1. Skip: `/mistake/i` or `/misinterpret/i` or `/confuse/i` or `/suspect/i` or `/worry/i` or `/surprise/i` or `/mimic/i` or `/simulate/i` or `/lesson/i` or `/\bhelp\b/i` or `/usually/i` or `/difficult/i` or `/pathart/i` or `/pathchallenge/i` or `/pathquiz/i` or `/pathgame/i` or `/^http/`
    - We skip tweets where (i) the pathologist discusses points of the case which may be easily mistaken – instead of providing a single diagnosis, (ii) the pathologist provides a diagnosis but may suspect an alternative diagnosis, or (iii) the tweet is simply a link to another tweet. No further keyword matching is performed for this tweet.
2. Tumoral status (ambiguously low grade or malignant): `/phyllod/i` or `/\bgrade\b/i` or `/tumou?r/i` or (`/[a-z]{3,}oma\b/i` and not `/schistost?oma/i`)

- Phyllodes tumors, mentions of “tumor” or “tumour”, mentions of tumor “grade”, and mentions of words that end in “oma” but are not “Schistosoma” – are all detected here.
- Loosely speaking, phyllodes tumors are only slightly more likely to be low grade than malignant. Because “tumor”, “-oma”, and “grade” do not necessarily mean a tumor is low grade or malignant, further keyword matching is performed. It is unlikely that the tweet is non-tumoral. If no other specific information is found after all further keyword matching is performed, the tumor is presumed to be low grade.
- *Schistosoma* (and its misspelling “Schistostoma”) refers to a genus of parasitic worm, rather than a tumor, though *Schistosoma* ends in “oma” like many tumor types.

3. Low grade: `/oma in situ/i` or `/chondroblastoma/i`

    - If we did not skip this tweet, but the tweet does mention “oma *in situ*”, e.g. “carcinoma *in situ*” or “melanoma *in situ*”, then we consider this tweet and images to represent low grade disease. Carcinoma *in situ* is pre-cancer, and we consider it more low grade than malignant. If a tweet contains only “carcinoma” but not “*in situ*”, subsequent steps will consider the tweet as malignant.
    - If the tweet includes “chondroblastoma”, this tweet is low grade. This is not to be confused with other blastomas, such as glioblastoma or lymphoblastoma, which are malignant and matched in subsequent steps.
    - No further keyword matching is performed if these patterns match. The tweet is low grade.

4. Malignant: `/malignant/i` or `/malignancy/i` or `/cancer/i` or `/\bCA\b/i` or `/carc?inoma/i` or `/sarcoma/i` or `/blastoma/i` or `/\bWilms/i` or `/GBM/i` or `/anaplas(?:ia|tic)/i` or `/metastas[ie]s/i` or `/metastatic/i` or `/\bmets?\b/i` or `/adenoca/i` or `/melanoma/i` or `/seminoma/i` or `/lymphoma/i` or `/leuka?emia/i` or `/mesothelioma/i` or `/myeloma/i` or `/hodgkin/i` or `/\bHL\b/i` or `/burkitt/i` or `/plasmoc[yi]toma/i` or (`/paget/i` and `/breast/i`) or `/\bCLL\b/i` or `/PCNSL/i` or `/NSCHL/i` or `/\bCHL\b/i` or `/NSCC/i` or `/\bI[LD]C\b/i` or `/\bASPS\b/i` or `/mtscc/i` or `/sq?cc/i` or `/rcc/i` or `/bcc/i`
   - Many diagnoses and abbreviations may indicate cancerous malignancy, e.g. carcinoma, sarcoma, Wilms' tumor, leukemia, RCC [renal cell carcinoma], NSCC [non-small cell lung carcinoma], or the standalone abbreviation "CA" [cancer].
   - We consider "anaplastic/anaplasia" to be more malignant than low grade disease.
   - No further keyword matching is performed if these patterns match. The tweet is malignant.
5. Nontumoral: `/congenital/i` or `/cholecystitis/i` or `/chorangiomatosis/i` or `/mycobacteri(?:um|al)\s*spindle\s*cell\s*pse?udotumor/i` or `/intravenous\s*leiomyomatosis/i` or `/helicobacter/i` or `/dirofilaria/i` or `/tuberculo/i` or `/enterobius/i` or `/echinococcus/i` or `/hydatid\s*cyst/i` or `/giardia/i` or `/cryptosporidium/i` or `/ascaris/i` or `/sarcina/i` or `/worm/i` or `/spiroquet(?:osis|es)/i` or `/diverticulosis/i` or `/villitis/i`

or `/colitis/i` or `/gastritis/i` or `/esophagitis/i` or `/appendicitis/i` or or `/xanthoma/i`

- Many diagnoses and abbreviations may indicate nontumoral disease, e.g. congenital conditions, *Helicobacter* infection, and villitis. Nontumoral disease keywords that contain "cyst", e.g. "cholecystitis" and "hydatid cyst", are detected here, because subsequent keyword matching steps will detect "cyst" as a sign of low grade tumoral disease.
- If one of these nontumoral keywords matches, no further keyword matches are attempted, and the tweet is considered nontumoral, even if prior steps detected "tumor" or "-oma". For instance, a "xanthoma" is a lipid aggregate, not a tumoral disease, even though xanthoma ends in -oma.

6. Low grade: `/benign/i` or `/cyst/i` or `/polyp/i` or `/hamartoma/i` or `/chorangioma/i` or `/ha?ematoma/i` or `/cylindroma/i` or `/fibroma/i` or `/luteoma/i` or `/c[yi]toma/i` or `/cond[yi]loma/i` or `/neoplas(?:ia|tic|m)/i` or `/LCIS/i` or `/DCIS/i` or `/\b[LD]IN\b/i` or `/lipoma/i` or `/carcinoid/i` or `/neuroma/i` or `/meningioma/i` or `/perineurioma/i` or `/cavernoma/i` or `/\bLGG\b/i` or `/\bODG\b/i` or `/oligodendroglioma/i` or `/craniopharyngioma/i` or `/le[yi]om[iy]oma/i` or `/schwannoma/i` or `/osteochondroma/i` or `/ependymoma/i` or `/angioma/i` or `/syringoma/i` or `/acanthoma/i` or `/collagenoma/i` or `/hidradenoma/i` or `/papilloma/i` or `/pilomatrixoma/i` or `/hydatidiform\s*mole/i` or `/wart/i` or `/molluscum/i` or `/\bHPV\b/i` or `/\bEBV\b/i` or `/kerat?osis/i` or `/fibrokeratoma/i` or `/melanoc[iy]tosis/i` or `/brenner/i`

or `/granular\s+cell\s+tumou?r/i` or `/metaplas(?:ia|tic)/i` or `/dysplas(?:ia|tic)/i` or `/dysembryoplas(?:ia|tic)/i` or `/hyperplas(?:ia|tic)/i` or `/\bLFH\b/i` or `/\bDNE?T\b/i` or `/\bNET\b/i` or `/\bPTC\b/i` or `/\bGIST\b/i` or `/\bSTIC\b/i` or `/\b[LD]ISN\b/i` or `/adenoma/i` or `/adenosis/i`

- Many diagnoses may indicate benign tumor, e.g. hamartoma, fibroma, condyloma, papilloma, lipoma, adenoma, adenosis, or cyst.
- We consider "neoplastic/neoplasia", "metaplastic/metaplasia", "hyperplastic/hyperplasia", and "dysplastic/dysplasia" to be more indicative of benign/low-grade/non-invasive/pre-malignant disease than malignant disease, but these terms are vague.
- We broadly consider oncovirus-driven tumors and wart-like growths to be in this low grade category also, e.g. HPV [human papilloma virus] warts and *Molluscum contagiosum* "water warts".
- We similarly consider abbreviations "LCIS" [lobular carcinoma *in situ*], "DCIS" [ductal carcinoma *in situ*], "LISN" [lobular *in situ* neoplasia], and "DISN" [ductal *in situ* neoplasia] to be more benign than malignant disease, so we categorize them as low grade. Though DCIS may require surgical or radiological intervention to be removed while LCIS may not, we consider our "low grade" and "malignant" categories to be defined by the apparent histopathology rather than the appropriate medical intervention. Predicting appropriate medical intervention would be a different machine learning task.
- If one of these keywords match, the tweet is considered low grade and no further keyword matching is performed.

7. Nontumoral: `/normal/i` or `/ulcer/i` or `/embolism/i` or `/thromb/i`

or `/rupture/i` or `/infarct/i` or `/aneurysm/i` or `/ha?emorrhag/i` or `/injur(?:y|ed)/i` or `/inflam/i` or `/swell/i` or `/balloon\s*cell\s*na?ev(?:us|i)/i` or `/decidua/i` or `/foreign/i` or `/lymphadenopath?y/i` or `/vasculopathy/i` or `/vasculitis/i` or `/synovitis/i` or `/pulmonary\s*interstitial\s*glycogenosis/i` or `/essential\s*thrombocythemia/i` or `/endometriosis/i` or `/mastoc[iy]tosis/i` or `/castleman/i` or `/herpe(?:s|tic)/i` or `/\bHSV\b/i` or `/\bCMV\b/i` or `/cytomegalovir/i` or `/viral/i` or `/bacteri(?:a|um)/i` or `/fung(?:al|us)/i` or `/mycetoma/i` or `/myco(?:sis|tic)/i` or `/infect(?:ion|ed)/i` or `/tauopathy/i` or `/amyloidosis/i` or `/neurodegen/i` or `/\brabies\b/i` or `/hemosiderosis/i` or `/polymicrogyria/i` or `/status\s*verrucosus/i` or `/\bIUGR\b/i` or `/storage\s*dis(?:ease|order)/i` or `/athero(?:sis|ma)/i` or `/atherosclero(?:sis|tic)/i` or `/gauzoma/i` or `/colchicine/i` or `/\bIBD\b/i` or `/GVHD/i` or `/crohn/i`

- Many diagnoses may indicate normal tissue of nontumoral disease, e.g. normal, embolism, decidua, tauopathy, foreign body, mycetoma, CMV [cytomegaolovirus] infection, GVHD [graft versus host disease], and Crohn's disease.
- If one of these nontumoral keywords matches, no further keyword matches are attempted, and the tweet is considered nontumoral, even if prior steps detected "tumor" or "-oma". For instance, a mycetoma is not a tumor, even though mycetoma ends with -oma.

8. Nontumoral: (not tumor/oma) and (`/#[a-z]*path/i` or

`/cerebell(?:um|ar)/i` or `/nodul(?:e|arity)/i`). Low grade if tumor/oma.

- If the tweet does not have tumor or "-oma" keywords detected from prior steps, and if the tweet has a #ANYpath hashtag (e.g. "#pulmpath" or "#pathology"), mention of "nodule"/"nodularity", or mention of the cerebellum, then we consider the tweet to be nontumoral. If instead the tweet has tumor or -oma keywords, then we consider the tweet to be low grade. The tweet is skipped if no steps identified the tweet as nontumoral, low grade, or malignant.
- Cerebellum is mentioned in several tweets, e.g. to depict normal cerebellar tissue[5]. Currently, we group normal tissue with tissue having nontumoral disease. We expect more tissue-based keywords may be used here in the future, as we expand our study to include more pathologists, tissues, and normal cases.
- In practice, we manually inspect all tweet message text to minimize the number of cases that are classified as nontumoral here. We typically write regular expressions to match specific keywords that indicate if a tweet represents nontumoral, low grade, or malignant disease.
- As part of our manual data curation, if on Twitter there was discussion among pathologists, and a different pathologist mentioned a correct diagnosis, and our consenting contributing pathologist concurred, then we write an auxiliary annotation file for the tweet with a summarized diagnosis[6]. This summary is also used for pattern

---

[5] Normal cerebellum case by S.Y. at `https://twitter.com/Sty_md/status/821840894634565632`

[6] A case of this is from author *K.H.*, where a different pathologist gave the diagnosis, and he

matching. This is an additional way that we minimize how many cases are handled at this late step.

- Moreover, if the contributing pathologist wrote diagnostic text directly in the image, we will write this text in the auxiliary annotation file for text matching also.[7]
- The way this "default nontumoral or low grade" rule is intended to be used is as a catch-all for unusual but non-malignant conditions[8]. Our motivation for this rule is to minimize our manual data curation burden. We do not wish to write an auxiliary annotation file or make a new regular expression for each unusual type of case, and we observe many of these cases are not malignant. It remains important to inspect the cases manually for correctness.

Tweets that do not match any nontumoral, low grade, or malignant rules are skipped in the same manner that Tweets matching skip rules are skipped. An additional caveat is this keyword matching may need refinement as we accumulate data.

### S4.5.8 Image features for machine learning

To perform baseline machine learning analyses on the images from social media, we derive a feature representation for each image, as follows. We crop each

---

agreed. We summarized this as "metastatic lobular carcinoma" in the auxiliary annotation file for the tweet `https://twitter.com/Ho_Khanh_MD/status/999989201734197250`.

[7] A case of this is from author *M.P.P.*, where M.P.P. wrote "IDC DIN LISN" directly on a shared histology image in the tweet `https://twitter.com/dr_MPrieto/status/890118713155997696` so we wrote this text in the auxiliary annotation file for the tweet.

[8] A case of this is from K.H., observing iron pill lesions in stomach biopsy `https://twitter.com/Ho_Khanh_MD/status/963800933716123648`.

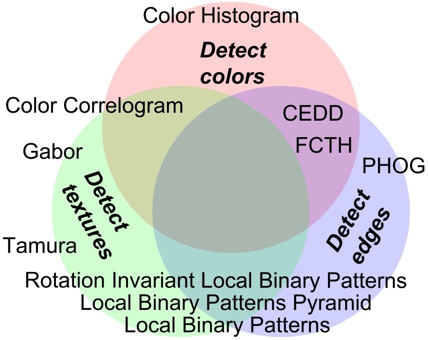

Color Histogram
Detect colors
Color Correlogram
CEDD
FCTH
Gabor
PHOG
Detect textures
Detect edges
Tamura
Rotation Invariant Local Binary Patterns
Local Binary Patterns Pyramid
Local Binary Patterns

Figure S4.19: **Machine learning features.** We use a variety of color, texture, and edge features for baseline machine learning analyses. Some features, such as color histograms, detect only color. Other features, such as Color Correlograms, detect both colors and textures. Pyramid features are scale-invariant. We separately consider SIFT features, which detect edges in a scale-invariant, rotation-invariant, and color-invariant manner, localized at interest points in an image.

image to the center square and resize it to 512×512 pixels [px]. See Sec S4.5.2 for more discussion of the 512×512px image size and how it relates to the 256×256px image size for the "overdrawn" criterion. This 512×512px image is then converted to a feature vector of 2,412 dimensions. The features we use (Fig S4.19) are available in Apache LiRE[119]. These features, and their dimension counts, are as follows: CEDD (144)[120], Color Correlogram (256)[121], Color Histogram (64)[119], FCTH (192)[122], Gabor (60)[119], Local Binary Patterns (256)[123], Local Binary Patterns Pyramid (756)[124], PHOG (630)[125], Rotation Invariant Local Binary Patterns (36)[126], and Tamura (18)[127].

### Deep learning instance and set feature vectors

After training, our appended 100-neuron layer (Fig 4.3B) is a 100-dimensional disease feature representation for a 224×224px patch. Due to our custom activation function and regularization (Fig S4.23), these 100 features are approximately binary (Fig S4.24C). A vector sum of these 100-dimensional approximately binary feature vectors is a 100-dimensional feature counter vector, which we relate to set cardinality (Eqn 4.8). Inspired by Deep Sets[128] and continuous bag-of-words[129] methods, we sample 21 patches throughout the white-balanced image, then add the 21 100-dimensional feature vectors to derive a 100-dimensional set representation for the white-balanced image (Fig 4.3C). In the set representation, if a feature's value is 21, then all 21 image patches have this feature. An approximate intuition follows that if the value is 20, then 1 of the 21 patches does not have this feature. This set-based approach offers limited interpretability and facilitates learning on large pathology images on social media, despite the receptive field of a deep neural network being much smaller (1 million pixels versus 224×224px, respectively). We train a Random Forest on the concatenation of (i) this 100-dimensional set representation, (ii) our 2,412-dimensional hand-engineered feature vector, (iii) the 10-dimensional tissue type covariate vector, and (iv) the 1-dimensional marker mention covariate vector (Fig 4.3C). Like Deep Sets, we use deep learning for instance learning, and add instance representations for a set representation. However, our approach differs from Deep Sets in that (i) we use a Random Forest to learn on set representations and side information (which is not differentiable end-to-end), (ii) we add approximately binary Centered Soft Clipping (Fig S4.23) features in the range (0,1) to implement counter-like set representations rather than add ReLU features in the range $[0,\infty)$ which do not necessarily count instance features, and (iii) we use Random Forest feature importance to interpret the relative influence of deep features, hand-engineered features, and optional covariates on prediction/classification (Fig 4.4).

### S4.5.9 Machine learning sanity checking for search

#### Prediction uncertainty quantified with ensemble

For disease state predictions posted to social media by the bot, we use an ensemble of classifiers to quantify prediction uncertainty. This ensemble consists of the set of classifiers for leave-one-pathologist-out precision@k search testing, so the ensemble size is equal to the number of pathologists who contributed data. Leaving one pathologist out from each classifier's training ensures some variability between classifiers. One classifier makes one prediction per image. We use a Z-test to determine if a disease state prediction's mean is significantly above chance. If no prediction is above chance, then the predictions may be due to chance alone and ignored by a pathologist. Likewise, search results may be ignored. This is our first prospective sanity check, and only requires a statistical interpretation of predictions.

**Training data detection with ensemble** This leave-one-out ensemble provides an additional non-prospective sanity check – if only one classifier in the ensemble makes a prediction that is strongly different from all the other classifiers in the ensemble, it is possible that the bot was requested to make a prediction on a case that was in the training data. This one "outlier" classifier that makes the strongly different prediction is the classifier that left this pathologist's case out for training. The distribution of predictions from these classifiers is depicted in a boxplot posted to social media by our bot (Fig 4.1 *at lower left*). An outlier is indicated by a circle in the boxplot, and a circle in a strongly different direction that the boxplot's interquartile range may suggest the "outlier classifier sanity check" has been encountered.

### Classifier repurposed, so if prediction suspect, then search suspect

If the Z-test of our distribution of predictions indicates the evidence in favor of a particular disease state, e.g. nontumor, is not due to chance alone, but this is surprising to a pathologist's expectations, the pathologist may consider the prediction to be suspect, so search results may be suspect as well. This is our second prospective sanity check, which requires a pathologists's expert opinion.

**Whole-patient prediction and disagreement detection** Our method calculates the probability of a disease state per image, but given 1-4 images are in a tweet, we are left with two questions. First, what is the overall probability that this patient has a particular disease state, e.g. malignant? For this, we make the naïve statistical independence assumption of probabilities, multiply the prediction probabilities, and normalize them to sum to one. Second, what should we do when the prediction for one image differs from another for a given patient? In this case, the bot includes a warning in its tweet message and suggests mistakes are more likely, but we do not consider this a prospective sanity check, because images could indeed show different disease states.

### Deep learning prediction heatmaps comparable to pathologist expectations

Heatmaps from our deep learning can localize disease states within an image (Figs 4.5, S4.21), offering our third prospective sanity check to pathologists. If localization of predicted disease is suspect, then prediction may be suspect, and search may be suspect.

### S4.5.10 Machine learning interpretability for search

#### Hand-engineered feature interpretability

We use existing hand-engineered visual features extensively (Fig S4.19). *Image features for machine learning* (Sec S4.5.8) discusses the combination of color, texture, and edge features we use (Fig S4.19). All of have human-defined mathematical or algorithmic behavior written in software code. We know by definition a color histogram feature is invariant to rotation, because such a feature may simply be the sum of each pixel's red value in an image. Similarly, rotating an image should not change the diagnosis, so rotation invariance makes sense for disease state prediction. We also know properties of other features, such as the important Local Binary Patterns Pyramid (LBPP)[124]. LBPP is globally color-invariant because it operates on grayscale pixel values, not color pixel values. This may provide robustness to staining protocol differences between institutions. LBPP is globally scale-invariant because it employs a pyramid for multi-scale representation. This is the same pyramid used by PHOG[125]. This pyramid may support robust machine learning despite pathologists sharing images at different magnifications. LBPP is locally rotation-invariant because it consists of rotation-invariant local binary patterns at every level of the multi-scale pyramid representation. These rotation-invariant local binary pattern features are locally robust to localized orientation changes of a pathology image, e.g. minor perturbations in the orientation of a cell or tissue fiber. In contrast, PHOG consists of oriented gradients at every pyramid level, rather than rotation-invariant local binary patterns. LBPP is not globally rotation-invariant because, like PHOG, most pyramid grid cells are spatially localized, so a rotated image will have a different feature representation. LBPP is a texture feature because it compares the value of a center pixel to the value of many pixels at a particular radius from the center. Prior groups have used texture features to distinguish stroma, lymphocytes, necrosis, etc. Hand-engineered feature interpretability provides a simple foundation on which to build more abstract levels of interpretability. We can also reason about what features do not improve disease state prediction, e.g. SIFT features, which are thought to cover nuclei.

### Random Forest feature importance

We use Random Forest feature importance to infer which features are important for disease state predictions. Hand-engineered visual features, clinical covariates, and deep learning features are concatenated together for a Random Forest to learn to predict disease state, so this single Random Forest classifier provides interpretability of each feature, in context together (Fig 4.4). We use this to infer broad principles, e.g. important clinico-visual features of disease state are texture (e.g. Local Binary Pattern Pyramid features) and tissue type covariates. Moreover, Random Forest feature importance identified several deep features were more important than the others for disease state prediction, so we focused our analyses on these important deep features.

For a before-and-after-histopathology-image-training comparison, we also consider feature importances when training a Random Forest with $\text{ImageNet}_{2048}$ features (Fig S4.20). $\text{ImageNet}_{2048}$ deep features have not been trained on histopathology images or the tissue type covariate. We observe that before histopathology training, these 2048 deep features are complemented by scale-invariant color-invariant edge features (i.e. PHOG) and rotation-invariant color features (color histograms and Color Correlograms). This may suggest disease

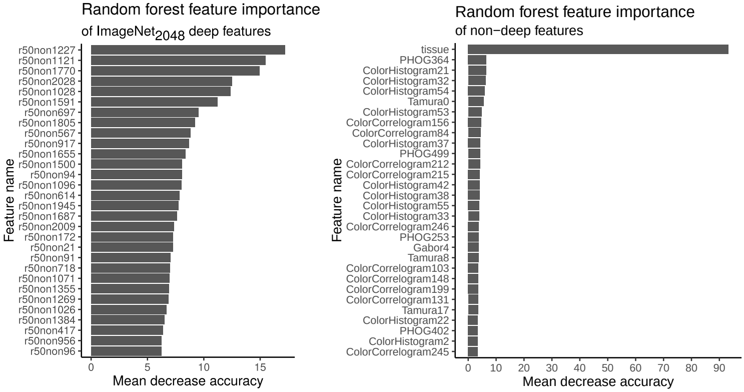


Figure S4.20: **Random Forest feature importance for prioritizing only-natural-image-trained deep features, when non-deep, ImageNet$_{2048}$, and clinical features are used together for learning.** As in Fig 4.4, we use the mean decrease in accuracy to measure Random Forest feature importance. However, we consider here ImageNet$_{2048}$ features, rather than the 100 deep features trained on histopathology images. No visual features here are designed with histopathology in mind or trained on histopathology data. This provides an interpretation of what hand-engineered and only-natural-image-trained deep features are important for disease state prediction, before training the deep neural network on histopathology images and covariates. In this way, we train a Random Forest on 2412 hand-engineered features, ImageNet$_{2048}$ features, and the tissue type covariate. The tissue type covariate is exceedingly important here, highlighting how disease state is reported in our data in a tissue-type-specific manner (Fig 4.8C1), e.g. infection is more likely reported in lung than breast. Pyramid histogram of oriented gradient (PHOG) features and color features (color histograms and Color Correlograms) are important hand-engineered features complimentary to the ImageNet$_{2048}$ features. PHOG is scale-invariant (due to pyramids) and color-invariant (due to grayscale), while said color histograms and Color Correlograms are rotation-invariant. Taken together, these important hand-engineered features may provide robust pathology representations for Random Forest learning. Such representations may complement ImageNet$_{2048}$ features, with "r50non1227" being the most important of these 2,048 features from the Global Average Pooling layer of a ResNet-50 trained only on the natural images (e.g. cats and dogs) of the ImageNet dataset. Thus of all 2,048 ImageNet$_{2048}$ features, r50non1227 may be prioritized first for interpretation (Fig S4.21 shows r50non1227 interpretation via heatmaps). *Random Forest feature importance* in the supplement discusses further (Sec S4.5.10).

state prediction benefits from these (i) invariant properties and (ii) features of color/stain intensity and distribution, that are not encoded in the $\text{ImageNet}_{2048}$ feature representation. Moreover, the tissue type covariate is strikingly important here (Fig S4.20). Therefore, after training the ResNet-50 on histopathology images and the tissue type covariate (Fig 4.3), we find (i) the tissue type covariate importance is reduced presumably because the ResNet-50 has to some extent learned to represent tissue type in its 100-dimensional feature vector, and (ii) scale-invariant and/or color-invariant texture features (e.g. Local Binary Pattern Pyramid [LBPP] and Local Binary Patterns [LBP]) become increasingly important presumably because the ResNet-50 has to some extent learned to represent pathology-relevant edges and color in its 100-dimensional feature vector while texture features are underrepresented. Thus texture features (i.e. LBPP/LBP), are important for disease state prediction, but the deep neural network did not learn similar texture features from the pathology data and learning methods at hand. We likewise reason that LBPP/LBP texture features may have low importance in the context of $\text{ImageNet}_{2048}$ features, because $\text{ImageNet}_{2048}$ features may represent similar texture, so LBPP/LBP are redundant with $\text{ImageNet}_{2048}$ for visual texture features predictive of disease state.

We note several of the most important $\text{ImageNet}_{2048}$ features (e.g. r50non1227, r50non1121, ...) have an importance measure (i.e. mean decrease in accuracy) greater than the most important hand-engineered features (e.g. PHOG364, ColorHistogram21, ...), which may suggest for disease state prediction that $\text{ImageNet}_{2048}$ features represent more predictive information than the best hand-engineered features we tested (Fig S4.20). Alternatively, the high importance of $\text{ImageNet}_{2048}$ features may be due to biases in Random Forest learning to choose features that take on many different values[130,98], as $\text{ImageNet}_{2048}$

features do.

In principle, for classification, any interpretable classifier may be used in place of the Random Forest, e.g. logistic regression, support vector machine, or generalized additive model[100]. A careful choice here may demonstrate favorable accuracy and interpretability. We choose a Random Forest as a simple baseline that requires (i) little tuning or preprocessing, (ii) learns interpretable nonlinear relationships among features and covariates, and (iii) provides a measure of similarity for search.

**Marker mention and SIFT features excluded from Random Forest feature importance analysis** We excluded from our Random Forest feature importance analysis the marker mention covariate and SIFT features, primarily because both did not improve 10-fold cross-validation prediction performance when using an ensemble of classifiers, which performed best for prediction (Fig 4.9: marker 0.8035±0.0043 vs 0.8025±0.0021, $U = 3$, $p = 0.7$; SIFT 0.8035±0.0043 vs 0.8014±0.0022, $U = 7$, $p = 0.4$, two-tailed Wilcoxon rank-sum test). Moreover, SIFT reduces performance when an ensemble is not used (0.7846±0.023 vs 0.7796±0.0019, $U = 85$, $p = 0.0004114$). This may suggest that for sufficiently strong disease state classifiers using H&E images and tissue covariates, the marker mention covariate and SIFT features provide at best only redundant information. For example, if the decision to order a marker test, e.g. IHC, is typically based on the H&E, and a classifier is sufficiently accurate at predicting disease state from H&E, the decision to order a marker test provides no additional disease state information. Secondarily, we excluded the marker mention covariate because it is based on the clinical opinion of all pathologists commenting on this case. Disease state is based on the diagnosis, which is also a clinical opinion. Rather than seeking to explain one opinion in terms of another opinion, we seek to explain opinions in terms of objective information in the H&E or clinicals, e.g. tissue type. We note 10-fold cross-validation may provide inflated measures of performance, so for a less inflated examination of the possible contributions of the marker covariate, SIFT features, and deep features, we turn to leave-one-pathologist-out cross-validation for search. *Disease state search, first pan-tissue pan-disease method* discusses this (Sec 4.3.6).

### Interpretability of important deep features through activation maps

Deep neural networks have a restricted field of view, but this is an advantage for interpretability, because one can systematically sweep a trained neural network across an image to localize deep feature activations. At each location in the sweep, a 224×224px image patch is fed to the trained neural network for interpretation. However, our Random Forest uses a set representation of the deep features, formed as a sum of deep feature vectors systematically sampled throughout the overall image of arbitrary size (Fig 4.1). Therefore, for spatial localization of disease state we make heatmaps to depict the deep feature activations at those sampled locations. These heatmaps indicate a correspondence between the most important deep features and the class labels. This approach facilitates deductive reasoning about predictions, e.g. (i) in Fig 4.5 the image overall is predicted by the deep-learning-random-forest hybrid classifier to be low grade (not shown), (ii) this classifier includes deep features (Fig 4.1C), (iii) a deep feature of ours is by definition a vector sum of images (Fig 4.1C) shown in the grid (Fig 4.5D2), (iv) the second most important deep feature (r50_30 in Fig 4.4) is known to correspond to low grade (e.g. Fig 4.5C2), (v) r50_30 is active

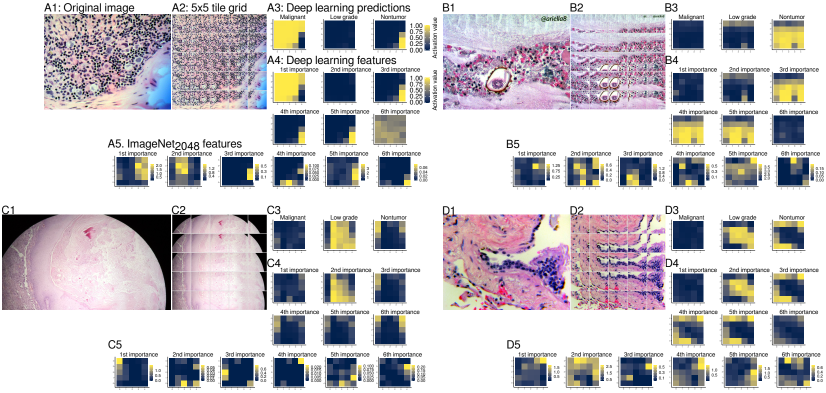

A1: Original image
A2: 5x5 tile grid
A3: Deep learning predictions
Malignant
Low grade
Nontumor
Activation value
A4: Deep learning features
1st importance
2nd importance
3rd importance
4th importance
5th importance
6th importance
A5. ImageNet2048 features
B1
B2
B3
B4
B5
@ariella8
C1
C2
C3
C4
C5
D1
D2
D3
D4
D5

Figure S4.21: **Interpretable spatial distribution of deep learning predictions and features of multiple cases.** We compare four cases side-by-side, in the manner of Fig 4.5. The most important histopathology-trained deep feature is r50_46 (Fig 4.4 *at left*), which predicts the majority class, e.g. **A3** *left* corresponds to **A4** *upper left*. The second most important feature (r50_30) predicts the second most abundant class, low grade. The third most important feature (r50_85) predicts the third most abundant of three classes, nontumor. The strong class correspondence in important deep features may suggest that removing the top layer's three class-predictor neurons may incur only a small loss of learned information (Fig 4.3B *top*). Shown in **A5** *at left* is the most important only-natural-image-trained deep feature (i.e. ImageNet$_{2048}$), "r50non1227" (Fig S4.20). Continuing left-to-right in **A5** we show ImageNet$_{2048}$ features in decreasing order of importance: r50non1121 (2nd more important), r50non1170 (3rd), r50non2028 (4th), r50non1028 (5th), and r50non1591 (6th). As expected, we find no intelligible pathology-related interpretation of these ImageNet$_{2048}$ features in these heatmaps, because these features are not trained on histopathology data. **(A)** *B.X.*: metastatic lobular carcinoma in satellite lymph node, where malignant activation is high throughout, except the lower right background. **(B)** *C.S.*: juvenile polyp, where nontumor activation is high both for the lentil at lower center (specifically rows 0-2 of columns 1-2 of the 5x5 grid of panel B2, where the lower left corner is row 0 of column 0) and the dark *Ascaris* ova at right (specifically the dark cluster in row 1 of column 4, and evident to some extent in rows 0-2 of columns 3-4), showing the breadth of the nontumor disease state category. **(C)** *R.S.*: is a proliferating epidermoid cyst. Despite viral wart change, we consider this in the low grade disease state (Fig S4.18). This example also illustrates a different image size and microscope eyepiece field of view artifact. **(D)** *Y.R.*: pulmonary vein lined by enlarged hyperplastic cells, which we consider to be low grade disease state, and **D4** *center top* highlights these low grade cells.

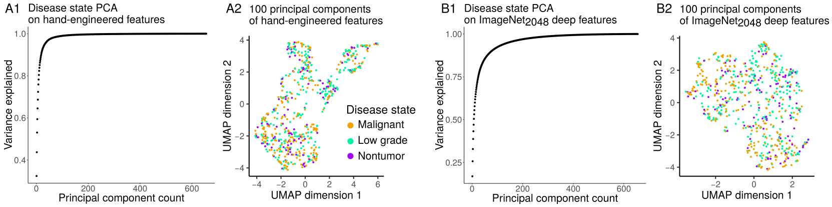


Figure S4.22: **Disease state clusters after dimensionality reduction.** As in Fig 4.6 we apply the UMAP[101] algorithm to determine if there are clusters of patient cases that have meaningful groups of features for the prediction of disease state. However, to investigate if cluster quality can be improved through dimensionality reduction, we first apply principal components analysis (PCA) to reduce hand-engineered feature dimensionality from 2,412 to 100 principal components, and follow the same procedure for ImageNet$_{2048}$ features. In practice, PCA is a common preprocessing step for the t-SNE clustering algorithm[131], but UMAP claims to have no computational restrictions on input dimension (so PCA is not expected to be required for UMAP)[101]. **(A1)** We show that 100 principal components explain 98.98% of the variance of the 2,412 hand-engineered features. Our histopathology-trained deep features are similarly 100 dimensions (Figs 4.3C, 4.6C1). **(A2)** As expected, PCA preprocessing does not qualitatively change UMAP clusters based on hand-engineered features. **(B1)** We show that 100 principal components explain 92.35% of the variance of the only-natural-image-trained ImageNet$_{2048}$ deep features. **(B2)** As expected, PCA preprocessing does not qualitatively change UMAP clusters based on ImageNet$_{2048}$ deep features. We conclude the vague clusters from hand-engineered features or lack of clusters from ImageNet$_{2048}$ is not a UMAP-related artifact of their high dimensionality, but instead simply means these features do not clearly group patients by disease state.

with a value of more than 0.5 for images shown in the grid center (Fig 4.5D2), (vi) therefore the classifier predicts low grade partly because images near the grid center have the low grade feature. Similar to our work, previous work used Random Forest feature importance for feature selection on a pretrained deep neural network[132], though to the best of our knowledge we are the first to use Random Forest feature importance for feature selection on a deep neural network retrained on the same task as the Random Forest.

### S4.5.11 Machine learning methods discussion

Because our image feature vectors are so wide, e.g. 2,412 dimensions (Fig S4.19) or more, we found best results with Random Forests when the number of features to consider for a decision/split was half (rounded up) of the total attribute count. This was especially important for covariates, e.g. the tissue covariate for disease state prediction. For search using Random Forest similarity, we grew each tree to a maximum depth of 10.

#### Deep learning

**Cross entropy loss for learning** We optimize the deep neural network with an unweighted cross entropy loss (Eqn 4.1, where $I_{y_i \in C_i}$ is the indicator function being 1 when the example $x$ class label $y_i$ is class $C_i$, and 0 otherwise)[9], for mini-batches of size $N = 64$ 224×224px images and typically $C = 3$ classes (nontumor, low grade, malignant). We use stratified bootstrap sampling for each epoch, so all classes have equal weight.

$$\mathcal{L}_{unweighted}(\{x_1, x_2, ..., x_N\}) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C}[I_{y_i \in C_i}]log(p_{model}[y_i \in C_i]) \qquad (4.1)$$

**ResNet-50 learning** For our deep learning, we freeze no layers of the ResNet-50. We train end-to-end with learning rate of 0.01 and Nesterov accelerated gradient momentum of 0.9[133,134,135]. We use Keras' learning rate decay of $10^{-5}$, which reduces the learning rate each batch. Moreover, we follow a learning rate schedule, where learning rate is divided by 100 until the end of epoch 1, but by 10 until the end of epoch 3. We train with Mixup[97], mixing according to draws

---

[9] For this formula please see `https://github.com/keras-team/keras/issues/6444`

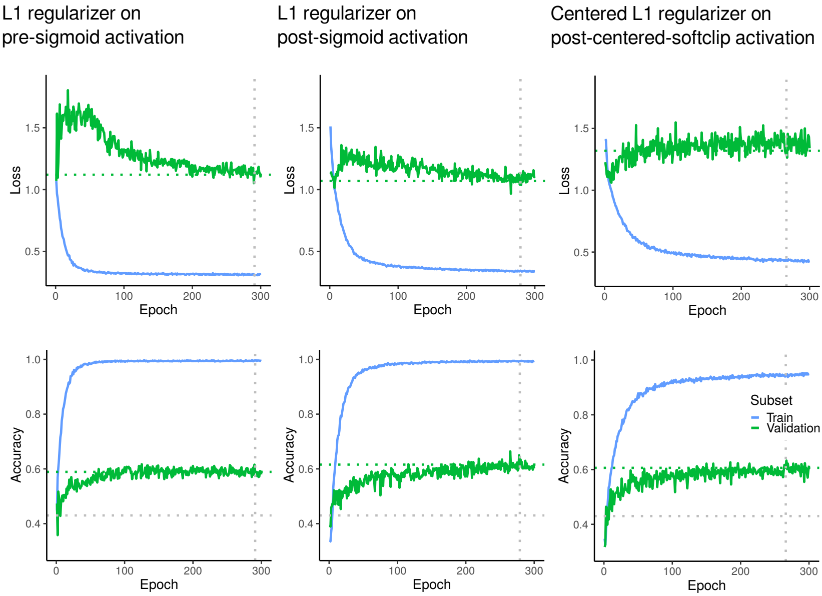


Figure S4.23: **Deep learning performance curves.** Loss and accuracy for training and validation sets, for one fold of 10-fold cross-validation, using a ResNet-50 to predict disease state, for individual 224×224px image patches. Validation accuracy improves with training, though high validation accuracy may suggest (i) the ground truth label may not apply universally throughout the image which is larger than 512×512px, (ii) samples whose state may not fit with our three-class disease state schema, (iii) mislabeled samples, (iv) deep learning overfit. Indeed, we see in many of our images mixtures of diseases and disease-free tissue, so (i) may be especially likely, and more advanced methods, such as multiple instance learning, may overcome this as we acquire more data. Observing that validation accuracy improved with training, we proceeded with this simple supervised deep learning approach as a proof-of-principle.

from a beta distribution with *alpha* = 1.4 and *beta* = 0.4. Our data augmentation is random flips, random free rotations, grayscale Gaussian noise in RGB color space (mean=0, stdev=0.5), and random brightness adjustment (uniform distribution, -0.05 to 0.05). We white-balance images before processing.

Freezing layers lowered validation accuracy. Alternative architectures such as DenseNet, Inception, and Xception trained more slowly and lowered validation accuracy. We do not report these results.

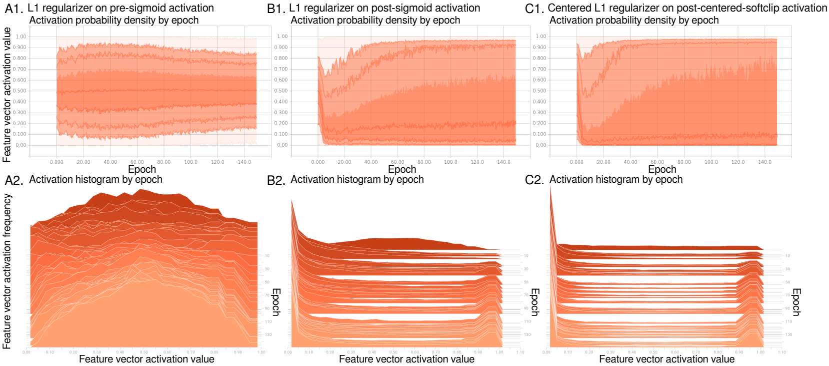


Figure S4.24: **Deep features.** We appended a 100 neuron fully connected layer on top of a ResNet-50 to learn a concise 100-dimensional "bottleneck" feature vector representation of a 224×224px image patch (Fig 4.3). This feature vector takes different distributions, depending on regularization. The top row shows a 100-dimensional activation distribution as standard deviations from the mean, and training proceeds left to right. The bottom row shows the same activation distribution as the top row, but as a histogram, with training proceeding from dark red at the top to orange at the bottom. **(A)** Feature values are centered at 0.5 when L1 regularization on the activation value before a sigmoid squashing function is applied, because a sigmoid transforms a 0 value to 0.5. **(B)** Feature values are approximately bimodal when L1 regularization on the activation value after sigmoid squashing is applied, because L1 regularization enforces sparsity by penalizing non-zero values. **(C)** Feature values are more strongly bimodal when our binarizing regularization (a.k.a. centered L1 regularization, Fig S4.25C3,D) on the activation value after our centered soft clipping activation function is applied, because this regularization penalizes values near 0.5 and centered soft clipping saturates to 0 or 1 more quickly than a sigmoid. We ultimately chose our centered L1 regularization and centered soft clipping activation to represent the deep feature space for disease state prediction, because (i) it performed better with Random Forest learning over feature vector sums of a set of multiple images (Fig S4.25A), (ii) it demonstrated comparable validation accuracy in deep learning (Fig S4.23), (iii) Random Forests have a bias to select features that take many values[130,98] so approximately binary deep features may reduce this bias by having a restricted distribution of values, and (iv) we believe approximately binary deep features to be simple and interpretable as set scalar cardinality, i.e. counters (Fig 4.5 and Eqn 4.8).

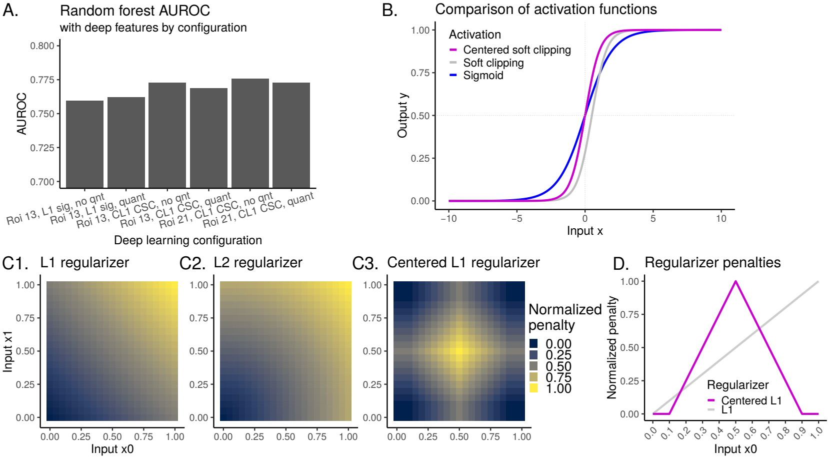


Figure S4.25: **Sigmoid vs centered soft clipping performance.** **(A)** Random Forest AUROC when ResNet-50 trained with sigmoid activation function with L1 regularization, versus centered soft clipping activation function with centered L1 regularization. Centered soft clipping demonstrates higher performance. Increasing the number of regions of interest (ROI) images in a deep set from 13 to 21 (Fig 4.1C shows 21) improves performance slightly. Quantizing features to either 0 or 1 slightly reduces performance when using centered soft clipping. **(B)** Comparison of sigmoid activation function with soft clipping[136] (Eqn 4.2) and our centered soft clipping (Eqn 4.3). Our centered soft clipping is centered at x=0 and steeper than sigmoid, which we argue is amenable for learning interpretable (Fig 4.5) and sharply binary hash codes. **(C)** Comparison of L1 (Eqn 4.11), L2, and our centered L1 (Eqn 4.13) regularizers in two dimensions, $x_0$ and $x_1$. Our L1 regularizer penalizes values close to 0.5, to encourage a binarized feature representation. **(D)** Comparison of L1 and our centered L1 regularizers, in one dimension, for clearer depiction of our centered L1 regularizer's margin $m = 0.1$ (Eqn 4.13), for stable learning and vanishing gradient avoidance.

### Deep set learning feature interpretation

As our ResNet-50 deep neural network trains (Fig S4.23), a 100-dimensional feature representation is learned (Fig S4.24), by the 100-neuron layer we append to the ResNet-50 (Fig 4.3C). A spatially-localized empirical interpretation of deep learning predictions and feature activations is available in Figure 4.5. We analytically interpret our deep features through their activation functions, regularization, and relationship with set scalar cardinality, below.

**Centered soft clipping activation function definition** To learn sharply binary features from the deep learning, we define a steep activation function, called centered soft clipping (CSC) (Fig S4.25B and Eqn 4.3). This is derived from soft clipping (SC)[136], which is not centered at x=0 (Fig S4.25B and Eqn 4.2). Like centered soft clipping, the sigmoid and hyperbolic tangent activation functions centered at x=0, but they are not as steep. There is a steepness parameter ($p$) in [centered] soft clipping, and we let $p = 2$ for our experiments (Eqns 4.2, 4.3).

$$\text{SC}_p(x) = \frac{1}{p} log(\frac{1 + e^{px}}{1 + e^{p[x-1]}}) \tag{4.2}$$

$$\text{CSC}_p(x) = \frac{1}{p} log(\frac{1 + e^{p[x+0.5]}}{1 + e^{p[x-0.5]}}) \tag{4.3}$$

**Centered soft clipping in the limit as Heaviside step function** Furthermore, we note in the limit $p \to \infty$, centered soft clipping converges to the Heaviside step function $H(x)$, a type of indicator function which is 1 for positive values, 0 for negative values, and 0.5 for zero values (Eqn 4.6). In the limit $p \to \infty$, $\text{CSC}_p(x)$ is a binary indicator of the presence or absence of a feature in a 224×224px image patch (Fig 4.1B), representing binary logic. Smaller values, e.g. $p = 2$, allow the representation of a small amount of probabilistic uncertainty regarding this presence or absence, where this area of uncertainty is infinitesimally small for $p \to \infty$.

$$\lim_{p\to\infty} \mathrm{CSC}_p(x) = \frac{1}{p}[log(1 + e^{p[x+0.5]}) - log(1 + e^{p[x-0.5]})] \tag{4.4}$$

$$\lim_{p\to\infty} \mathrm{CSC}_p(x) = \begin{cases} \frac{1}{p}[log(1+0) - log(1+0)] = \frac{0}{p} = 0 & \forall x < 0, \\ \frac{1}{p}[log(1+e^{p(0.5)}) - log(1+0)] = \frac{p(0.5)}{p} = 0.5 & \forall x = 0, \\ \frac{1}{p}[p(x+0.5) - p(x-0.5)] = \frac{p}{p} = 1 & \forall x > 0 \end{cases} \tag{4.5}$$

$$\lim_{p\to\infty} \mathrm{CSC}_p(x) = H(x) = \begin{cases} 0 & \forall x < 0, \\ 0.5 & \forall x = 0, \\ 1 & \forall x > 0 \end{cases} \tag{4.6}$$

**Deep feature as set scalar cardinality** To our Random Forest, a deep feature is the scalar cardinality (Eqns 4.7-4.8) of the set of these presences measured at 21 different locations throughout the original image (Fig 4.1C). Here, $x$ represents one of the 100 features from the trained layer atop the ResNet-50 in Figure 4.1B, and $I$ is an indicator function of set membership.

$$\mathrm{Card}(\mathrm{set}) = |\mathrm{set}| = |\{x_1, x_2, \ldots, x_N\}| = \sum_{n=1}^{N=21} I(x_n) \tag{4.7}$$

$$= \sum_n H(x_n) = \sum_n [\lim_{p\to\infty} \mathrm{CSC}_p(x_n)] \approx \sum_n \mathrm{CSC}_p(x_n) \;\; \forall p > 0 \tag{4.8}$$

**Data as Bernoulli process, deep features, and expected value** To consider some input data passed through the ResNet-50, we model a random set as a Bernoulli process, with each set member taking some value on the interval $(0, \infty)$ at probability $\varphi$, otherwise some value on the interval $(-\infty, 0)$. Let $\mathbb{P}_\varphi(x)$ be a Bernoulli trial that is $(0, \infty)$ at probability $\varphi$, otherwise $(-\infty, 0)$ A Heaviside step function (Eqn 4.6), which our centering soft clipping (CSC) activation function approximates in the limit of $p \to \infty$, transforms this $\mathbb{P}_\varphi(x)$ to a standard Bernoulli trial taking value 1 at probability $\varphi$, otherwise 0. Thus the cardinality of this random set is proportional to the Bernoulli trial's expected value $\mathbb{E} = \varphi$ (Eqn 4.10). In a sense, for a fixed set size e.g. $N = 21$, the Random Forest may learn what thresholds of $\mathbb{E}$ of these deep features are predictive of the classes of interest, in the context of the other available deep and non-deep features (Fig 4.1C).

$$\mathrm{Card}(\mathrm{set}) = |\mathrm{set}| = |\{\mathbb{P}_\varphi(x_1), \mathbb{P}_\varphi(x_2), \ldots, \mathbb{P}_\varphi(x_N)\}| = \sum_n^N H(\mathbb{P}_\varphi(x_n)) = \tag{4.9}$$

$$= \sum_n^N [\lim_{p\to\infty} \mathrm{CSC}_p(\mathbb{P}_\varphi(x_n))] = N\mathbb{E}(x) = N\varphi \tag{4.10}$$

**Centered L1 regularizer in the context of Gaussian initialization** Neural network weights are often initialized with Gaussian noise having mean of zero[133], so centering the activation function at zero removes initial bias towards learning a 0 or 1 when combined with our centered L1 regularizer (Eqn 4.3). Whereas a standard L1 regularizer (Eqn 4.11) penalizes all non-zero values, our centered L1 regularizer penalizes values close to 0.5 (Eqn 4.13), with the assumption that all values are between 0 and 1, which is true for values from centered soft clipping (Eqn 4.3).

$$\mathrm{L1}(x_0, x_1) = |x_0 + x_1| \tag{4.11}$$

$$\mathrm{CL1penalty}(x, m) = max(m, 0.5 - |x - 0.5|) - m \tag{4.12}$$

$$\mathrm{CL1}(x_0, x_1, m) = |\mathrm{CL1penalty}(x_0, m) + \mathrm{CL1penalty}(x_1, m)| \tag{4.13}$$

To approximately binarize deep features, our centered L1 regularizer penalizes output values close to 0.5, which occur when $x = 0$ with sigmoid and centered soft clipping activations. (Fig S4.25D). For stable learning, and to help avoid the vanishing gradient problem[137,138] from values infinitesimally close to 0 or 1, our centered L1 regularizer has no penalty for values less than 0.1 or greater than 0.9. Restated, we let the L1 regularizer's margin parameter $m \in [0, 0.5]$ be $m = 0.1$.

**Binary hash code learning** Lin *et al*[139]. use deep learning with sigmoid activations to learn binary hash codes for search of clothing images. We intend to learn sharply binary hash codes with centered soft clipping and centered L1 regularization together, for interpretability, pathology search, and set representation learning. We find centered soft clipping performed slightly better for classification (Fig S4.25A) and provided a small but significant increase in search performance (Fig 4.10B).

### S4.5.12 Disease state search distance function and performance

Pathology search performance is shown in Figure 4.10B, and detailed in Table S4.1. We found a combined distance measure (Eqn 4.14) worked best for search (i.e. Table S4.1 top row). The combined distance between two examples (i.e. $\mathrm{Dist}(x_0, x_1)$) is the number of trees in the Random Forest (i.e. 1000), minus the Random Forest similarity of the covariates and non-deep features (i.e. $\mathrm{Sim}_{\mathrm{RF}}(x_0, x_1)$), plus five times (i.e. $\alpha_{SIFT} = 5$) the L1 norm of the 5 largest SIFT cluster medoids vector sum (i.e. $\mathrm{L1}_{\mathrm{SIFT}_{\mathrm{k5}}}(x_0, x_1)$), plus the L1 norm of the top three deep features from each example (i.e. $\mathrm{L1}_{\mathrm{Deep}_3}(x_0, x_1)$). For $\mathrm{Deep}_3$, we use the top three most class-associated features of example $x$, one such feature for each of the three classes (nontumor, low grade, or malignant per Fig 4.5C). We initially found these three features using Random Forest similarity (Fig 4.4), but subsequently identify these features by minimizing the feature's error with

Table S4.1: **Case similarity search performance in detail.** Precision@k for search as shown in Figure 4.10B, with additional experiments. To estimate 95% confidence intervals, recall standard error of the mean is the sample standard deviation (i.e. stdev) divided by $\sqrt{10}$, since we perform ten replicates. This table and further details continue on the following page.

| Method († = *plotted in Fig 4.10B*) | prec@k=1 | prec@k=2 | prec@k=3 | prec@k=4 | prec@k=5 |
|---|---|---|---|---|---|
| | prec@k=6 | prec@k=7 | prec@k=8 | prec@k=9 | prec@k=10 |
| *RandomForest(2412 + tissue + marker) + 5×SIFT$_{k5}$ + Deep$_3$ + 0.01×ImageNet$_{2048}$†* | 0.7618±0.0018 | 0.7003±0.0016 | 0.6595±0.0011 | 0.6321±0.0005 | 0.6124±0.0008 |
| | 0.5977±0.0006 | 0.5846±0.0006 | 0.5739±0.0008 | 0.5649±0.0008 | 0.5571±0.0004 |
| RandomForest(2412 + tissue + marker) + 5×SIFT$_{k5}$ + Deep$_3$ + 0.001×ImageNet$_{2048}$ | 0.7567±0.0016 | 0.6988±0.0015 | 0.6597±0.0007 | 0.6319±0.0007 | 0.6129±0.0011 |
| | 0.5981±0.0011 | 0.5853±0.0007 | 0.5750±0.0006 | 0.5660±0.0009 | 0.5582±0.0008 |
| RandomForest(2412 + tissue + marker + ImageNet$_{2048}$) + 5×SIFT$_{k5}$ + Deep$_3$ | 0.7566±0.0014 | 0.6981±0.0010 | 0.6594±0.0006 | 0.6317±0.0010 | 0.6124±0.0012 |
| | 0.5974±0.0009 | 0.5850±0.0006 | 0.5748±0.0006 | 0.5658±0.0010 | 0.5580±0.0008 |
| RandomForest(ImageNet$_{2048}$ + tissue + marker) | 0.7517±0.0025 | 0.6940±0.0014 | 0.6563±0.0008 | 0.6294±0.0012 | 0.6113±0.0012 |
| | 0.5962±0.0008 | 0.5842±0.0010 | 0.5738±0.0007 | 0.5644±0.0006 | 0.5565±0.0006 |
| RandomForest(2412 + tissue + marker) + 5×SIFT$_{k5}$ + Deep$_3$ + 0.1×ImageNet$_{2048}$ | 0.7341±0.0012 | 0.6639±0.0010 | 0.6256±0.0009 | 0.5969±0.0005 | 0.5784±0.0004 |
| | 0.5620±0.0006 | 0.5492±0.0005 | 0.5381±0.0003 | 0.5296±0.0004 | 0.5220±0.0003 |
| *RandomForest(2412 + tissue + marker) + 5×SIFT$_{k5}$ + Deep$_3$†* | 0.7006±0.0026 | 0.6493±0.0023 | 0.6196±0.0017 | 0.5983±0.0009 | 0.5826±0.0007 |
| | 0.5698±0.0007 | 0.5597±0.0007 | 0.5511±0.0008 | 0.5440±0.0007 | 0.5380±0.0005 |
| RandomForest(2412 + tissue + marker) + 1×SIFT$_{k5}$ + Deep$_3$ | 0.6991±0.0023 | 0.6483±0.0022 | 0.6194±0.0012 | 0.5984±0.0009 | 0.5833±0.0008 |
| | 0.5708±0.0007 | 0.5610±0.0007 | 0.5519±0.0006 | 0.5443±0.0008 | 0.5382±0.0005 |
| RandomForest(2412 + tissue + marker) + Deep$_3$ | 0.6983±0.0016 | 0.6479±0.0018 | 0.6187±0.0017 | 0.5977±0.0012 | 0.5828±0.0009 |
| | 0.5705±0.0011 | 0.5604±0.0009 | 0.5516±0.0006 | 0.5442±0.0005 | 0.5381±0.0006 |
| RandomForest(2412 + tissue + marker) + 5×SIFT$_{k5}$ + EnsDeep$_3$ | 0.6982±0.0026 | 0.6491±0.0013 | 0.6177±0.0013 | 0.5965±0.0013 | 0.5811±0.0007 |
| | 0.5687±0.0007 | 0.5591±0.0011 | 0.5505±0.0010 | 0.5432±0.0010 | 0.5376±0.0011 |
| RandomForest(2412 + tissue + marker) + EnsDeep$_3$ | 0.6974±0.0025 | 0.6474±0.0016 | 0.6182±0.0015 | 0.5970±0.0014 | 0.5815±0.0007 |
| | 0.5689±0.0009 | 0.5594±0.0007 | 0.5510±0.0007 | 0.5440±0.0008 | 0.5381±0.0007 |
| RandomForest(2412 + tissue + marker) + 5×SIFT$_{k5}$ | 0.6948±0.0032 | 0.6449±0.0021 | 0.6154±0.0016 | 0.5941±0.0016 | 0.5791±0.0014 |
| | 0.5668±0.0011 | 0.5569±0.0011 | 0.5479±0.0008 | 0.5404±0.0002 | 0.5344±0.0007 |
| RandomForest(2412 + tissue + marker) + 1×SIFT$_{k5}$ | 0.6935±0.0029 | 0.6439±0.0021 | 0.6152±0.0012 | 0.5943±0.0016 | 0.5793±0.0015 |
| | 0.5669±0.0012 | 0.5569±0.0012 | 0.5484±0.0009 | 0.5412±0.0008 | 0.5350±0.0008 |
| RandomForest(2412 + tissue + marker + SIFT$_{k5}$) + Deep$_3$ | 0.6918±0.0030 | 0.6436±0.0021 | 0.6166±0.0020 | 0.5965±0.0017 | 0.5819±0.0010 |
| | 0.5699±0.0010 | 0.5608±0.0009 | 0.5518±0.0008 | 0.5443±0.0008 | 0.5380±0.0008 |
| *RandomForest(2412 + tissue + marker)†* | 0.6908±0.0021 | 0.6435±0.0017 | 0.6148±0.0015 | 0.5941±0.0013 | 0.5793±0.0016 |
| | 0.5671±0.0014 | 0.5568±0.0009 | 0.5482±0.0009 | 0.5409±0.0012 | 0.5351±0.0009 |
| RandomForest(2412 + tissue) + Deep$_3$ | 0.6602±0.0022 | 0.6129±0.0019 | 0.5845±0.0014 | 0.5656±0.0010 | 0.5530±0.0005 |
| | 0.5419±0.0008 | 0.5333±0.0006 | 0.5265±0.0010 | 0.5209±0.0010 | 0.5156±0.0005 |
| *RandomForest(2412 + tissue)†* | 0.6533±0.0025 | 0.6064±0.0023 | 0.5784±0.0013 | 0.5599±0.0010 | 0.5481±0.0009 |
| | 0.5386±0.0008 | 0.5297±0.0009 | 0.5227±0.0008 | 0.5167±0.0007 | 0.5119±0.0007 |
| RandomForest(2412 + tissue + SIFT$_{k5}$) + Deep$_3$ | 0.6527±0.0024 | 0.6073±0.0027 | 0.5803±0.0014 | 0.5638±0.0018 | 0.5518±0.0016 |
| | 0.5412±0.0016 | 0.5331±0.0012 | 0.5262±0.0010 | 0.5202±0.0012 | 0.5148±0.0010 |
| L1(ImageNet$_{2048}$) | 0.6432±0.0000 | 0.5839±0.0000 | 0.5543±0.0000 | 0.5346±0.0000 | 0.5204±0.0000 |
| | 0.5099±0.0000 | 0.5002±0.0000 | 0.4930±0.0000 | 0.4874±0.0000 | 0.4830±0.0000 |
| RandomForest(ImageNet$_{2048}$) | 0.6406±0.0046 | 0.5875±0.0028 | 0.5572±0.0021 | 0.5390±0.0017 | 0.5253±0.0019 |
| | 0.5144±0.0013 | 0.5057±0.0009 | 0.4988±0.0008 | 0.4934±0.0008 | 0.4887±0.0007 |
| RandomForest(2412 + tissue + marker) + SIFT | 0.6376±0.0014 | 0.5875±0.0016 | 0.5549±0.0010 | 0.5351±0.0009 | 0.5214±0.0009 |
| | 0.5119±0.0005 | 0.5034±0.0008 | 0.4961±0.0006 | 0.4900±0.0004 | 0.4851±0.0005 |
| *RandomForest(2412 + tissue + marker + SIFT)†* | 0.5728±0.0022 | 0.5329±0.0023 | 0.5108±0.0013 | 0.4969±0.0009 | 0.4874±0.0010 |
| | 0.4803±0.0011 | 0.4751±0.0008 | 0.4701±0.0007 | 0.4654±0.0008 | 0.4617±0.0007 |
| RandomForest(2412) + Deep$_3$ | 0.5720±0.0036 | 0.5363±0.0027 | 0.5153±0.0021 | 0.5014±0.0024 | 0.4915±0.0016 |
| | 0.4839±0.0015 | 0.4774±0.0009 | 0.4728±0.0010 | 0.4692±0.0008 | 0.4658±0.0007 |

Table S4.1: **Case similarity search performance in detail *(continued)*.** L1(. . . ) indicates distance is the L1 norm of the feature vector. RandomForest(. . . ) indicates distance is the number of trees in the forest (i.e. 1000) minus Random Forest similarity calculated on the feature vector. Tissue and/or marker covariate use are indicated by “tissue” and/or “marker”, respectively. “2412”/“2412 features” indicates our 2,412 hand-engineered features are used (Fig S4.19). However, “ImageNet$_{2048}$” indicates the 2,048 top-level features are used from a ResNet-50 trained only on ImageNet images. No histology images are used to train this ResNet-50’s 2,048 features, but these features are summed over 21 locations, just as the ‘histology-image-trained ‘Deep” features are (Fig 4.1C). “Deep” indicates the full 100-dimensional feature vector is used (Fig 4.1B,C). “Deep$_3$” indicates only the top 3 class-correlated features are used (Fig 4.5, see Eqn 4.14 for combining distances). “EnsDeep$_3$” indicates the top 3 class-correlated features are used, averaged across three neural networks in an ensemble. “SIFT” indicates all SIFT interest points are summed to represent an image. “SIFT$_{k5}$” indicates the 5 largest of 25 medoids are summed to represent an image. “5×SIFT$_{k5}$” indicates the SIFT$_{k5}$ vector is multiplied by five (i.e. $\alpha_{\text{SIFT}} = 5$), which changes L1-based distances (Eqn 4.14). We find best performance for “RandomForest(2412 + tissue + marker) + 5×SIFT$_{k5}$ + Deep$_3$”, which is 1000 minus Random Forest similarity, plus five times L1(SIFT$_{k5}$), plus L1(Deep$_3$).

| Method († = *plotted in Fig 4.10B*) | prec@k=1 | prec@k=2 | prec@k=3 | prec@k=4 | prec@k=5 |
|---|---|---|---|---|---|
| | prec@k=6 | prec@k=7 | prec@k=8 | prec@k=9 | prec@k=10 |
| *RandomForest(2412 features)†* | 0.5640±0.0024 | 0.5283±0.0025 | 0.5085±0.0018 | 0.4955±0.0021 | 0.4854±0.0014 |
| | 0.4775±0.0012 | 0.4717±0.0010 | 0.4667±0.0010 | 0.4631±0.0007 | 0.4600±0.0006 |
| *L1(2412 features)†* | 0.5479±0.0000 | 0.5124±0.0000 | 0.4927±0.0000 | 0.4831±0.0000 | 0.4746±0.0000 |
| | 0.4672±0.0000 | 0.4626±0.0000 | 0.4568±0.0000 | 0.4536±0.0000 | 0.4510±0.0000 |
| RandomForest(Deep + tissue + marker) | 0.5455±0.0020 | 0.5206±0.0018 | 0.5067±0.0013 | 0.4964±0.0006 | 0.4895±0.0010 |
| | 0.4839±0.0008 | 0.4796±0.0007 | 0.4758±0.0006 | 0.4727±0.0008 | 0.4701±0.0006 |
| RandomForest(2412 + tissue + Deep) | 0.5427±0.0025 | 0.5135±0.0015 | 0.5001±0.0015 | 0.4915±0.0017 | 0.4860±0.0011 |
| | 0.4806±0.0010 | 0.4756±0.0007 | 0.4719±0.0007 | 0.4690±0.0007 | 0.4671±0.0007 |
| *RandomForest(SIFT)†* | 0.4636±0.0024 | 0.4489±0.0021 | 0.4389±0.0016 | 0.4353±0.0013 | 0.4326±0.0009 |
| | 0.4304±0.0011 | 0.4282±0.0008 | 0.4266±0.0008 | 0.4249±0.0009 | 0.4235±0.0009 |
| *L1(SIFT)†* | 0.4630±0.0000 | 0.4495±0.0000 | 0.4399±0.0000 | 0.4376±0.0000 | 0.4324±0.0000 |
| | 0.4296±0.0000 | 0.4268±0.0000 | 0.4260±0.0000 | 0.4252±0.0000 | 0.4242±0.0000 |
| L1(SIFT$_{k5}$) | 0.4205±0.0000 | 0.4127±0.0000 | 0.4125±0.0000 | 0.4109±0.0000 | 0.4107±0.0000 |
| | 0.4107±0.0000 | 0.4102±0.0000 | 0.4104±0.0000 | 0.4112±0.0000 | 0.4105±0.0000 |
| RandomForest(SIFT$_{k5}$) | 0.4102±0.0050 | 0.4097±0.0020 | 0.4087±0.0017 | 0.4088±0.0014 | 0.4089±0.0013 |
| | 0.4092±0.0017 | 0.4090±0.0015 | 0.4088±0.0015 | 0.4087±0.0012 | 0.4088±0.0011 |
| *RandomForest(2412 + tissue),* | 0.3967±0.0044 | 0.3991±0.0036 | 0.3999±0.0033 | 0.3996±0.0030 | 0.4001±0.0033 |
| *permutation test†* | 0.3999±0.0027 | 0.4000±0.0029 | 0.3996±0.0030 | 0.3997±0.0027 | 0.3998±0.0026 |

respect to a class label in the training data, for computational expediency.

$$\mathrm{Dist}(x_0, x_1) = 1000 - \mathrm{Sim}_{\mathrm{RF}}(x_0, x_1) + [\alpha_{\mathrm{SIFT}}]\mathrm{L1}_{\mathrm{SIFT}_{\mathrm{k5}}}(x_0, x_1) + \mathrm{L1}_{\mathrm{Deep}_3}(x_0, x_1) \quad (4.14)$$

| | median | median 95% CI | mean | stdev | mean 95% CI |
|---|---|---|---|---|---|
| prec@k=1 | 0.7819 | 0.7499-0.8065 | 0.7470 | 0.1167 | 0.6978-0.7896 |
| prec@k=2 | 0.7237 | 0.6643-0.7482 | 0.6941 | 0.0967 | 0.6550-0.7300 |
| prec@k=3 | 0.6766 | 0.6320-0.6938 | 0.6523 | 0.1030 | 0.6115-0.6922 |
| prec@k=4 | 0.6373 | 0.6000-0.6769 | 0.6304 | 0.0975 | 0.5920-0.6676 |
| prec@k=5 | 0.6112 | 0.5771-0.6563 | 0.6123 | 0.0962 | 0.5742-0.6501 |
| prec@k=6 | 0.5982 | 0.5401-0.6451 | 0.5984 | 0.0931 | 0.5622-0.6345 |
| prec@k=7 | 0.5858 | 0.5236-0.6378 | 0.5847 | 0.0907 | 0.5494-0.6203 |
| prec@k=8 | 0.5736 | 0.5117-0.6301 | 0.5741 | 0.0879 | 0.5399-0.6079 |
| prec@k=9 | 0.5620 | 0.5042-0.6230 | 0.5644 | 0.0886 | 0.5296-0.6003 |
| prec@k=10 | 0.5526 | 0.4964-0.6132 | 0.5558 | 0.0884 | 0.5210-0.5899 |

Table S4.2: **Case similarity search performance per-pathologist.** Precision@k per-pathologist, corresponding to Fig 4.10C *at right*, for the best performing search method *RandomForest(2412 + tissue + marker) + 5×*$SIFT_{k5}$ *+* $Deep_3$ *+ 0.01×*$ImageNet_{2048}$, c.f. Table S4.1. Leave-one-pathologist-out cross-validation is repeated ten times. We considered 24 of 25 pathologists for this analysis, excluding one pathologist who shared no human acceptable H&E images for disease state prediction (Fig S4.14 details acceptability criteria). Each included pathologist's performance is averaged over these ten repetitions. This average performance is plotted in Fig 4.10C, where the 24 lines correspond to 24 pathologists, and error bars are standard error for these ten replicates. For precision@k=1 mean, each of the 24 pathologists' ten-repetition average precision@k=1 is averaged and reported. In this way, the mean prec@k=1 (i.e. 0.7470) is what precision@k=1 may be expected for a pathologist whose data we have not seen before, when averaged over many images for that pathologist. Bootstrapped confidence intervals (CI) for median and mean are shown for 10,000 replicates. There is not a significant difference between the means and medians, so any differences may be due to chance alone.

## S4.5.13 Supplementary Disease state search results

### Clinical covariates improve search performance

We ask if pathology-specific clinical covariates improve search performance. Including the tissue type covariate significantly improves performance compared to not using this covariate (0.5640±0.0024 vs 0.6533±0.0025, $U$ =

$100, p = 0.0001796$). We find including the marker mention covariate significantly improves performance further ($0.6533 \pm 0.0025$ vs $0.6908 \pm 0.0021$, $U = 100, p = 0.0001796$). Therefore including pathology-specific clinical covariates for pathology search is currently justified, because these improve search performance. We reason that disease states are reported at tissue-type-specific rates, and cases that mention marker tests (e.g. IHC) tend to be more similar to other cases that mention marker tests. Often, marker tests are used to subtype malignancies (e.g. TTF-1), but this is not always the case (e.g. Ki-67). We believe this is the first multimodal classifier that demonstrates improved search performance when combining pathology imaging features with clinical covariates (i.e. tissue type and mention of marker tests) that may be missing for some patients (Fig 4.10B).

### In the context of other features, nuclear features of disease are better represented by the most prevalent SIFT clusters rather than all SIFT

**Used alone, all SIFT is better than chance, but does not complement other features** We ask if cell nuclei features, as represented by SIFT, may represent disease state, and if so, which SIFT representations perform better than others for pathology search. Inspired by continuous bag-of-words methods that represent a context as a vector sum of words[129] and bag-of-visual-words methods that leverage SIFT[140], our simplest baseline detects all SIFT interest points in an image, then takes the sum of all these SIFT feature vectors to represent an image overall. This way, an image is represented by a 128-dimensional SIFT set representation, where the set cardinality is equal to the number of SIFT interest points in the image, which varies among images. Because SIFT features are non-negative and cover nuclei, the magnitude of this vector may increase as nuclear density increases. This performs better than chance when used alone (0.4636±0.0024 vs 0.3967±0.0044, $U = 100$, $p = 1.083e\text{-}05$), but reduces performance when used in the context of our 2,412 hand-engineered features, tissue covariate, and marker covariate (0.6908±0.0021 vs 0.5728±0.0022, $U = 100$, $p = 0.0001796$). We conclude that naïvely using SIFT features this way to estimate nuclear density allows limited discrimination of disease state when used alone, but does not complement other features when used together for pathology search.

**SIFT learning challenging, but a combined distance function of L1 norm and Random Forest similarity improves search** In light of the aforementioned reduced performance with SIFT, we ask if Random Forest similarity contributes to this performance reduction. We observe the SIFT L1 norm and SIFT Random Forest similarity performance differences may be due to chance alone (0.4630±0.0000 vs 0.4636±0.0024, $U = 40$, $p = 0.4429$). The Random Forest does not appear to learn from this SIFT representation. We additionally observe search performance is reduced less when adding the SIFT L1 norm to Random Forest similarity as a combined distance measure (Eqn 4.14), rather than training the Random Forest on SIFT features (0.5728±0.0022 vs 0.6376±0.0014, $U = 0$, $p = 0.0001796$). This may suggest the Random Forest overfits or cannot learn from these continuous bag-of-words SIFT features, but this may be mitigated by a combined distance function that does not rely on Random Forest learning from SIFT features (Eqn 4.14).

**SIFT clusters provide complimentary information for search** We then ask if SIFT may provide complimentary information for search, when using SIFT to estimate nuclear shape and edge distributions, rather than nuclear density. For this approach, we (i) detect all SIFT interest points in an image, (ii) form 25 clusters using k-medoids++ clustering with the L1 norm of the SIFT feature vectors for all interest points, (iii) retain the 5 medoids corresponding to the 5 most abundant clusters, and (iv) take the sum of the medoid SIFT features to represent the image overall. This way, an image is represented by a 128-dimensional SIFT set representation, where the set cardinality is always 5, the number of retained medoids. We call this a $\text{SIFT}_{\text{k5}}$ set representation, for k=5 medoids from k-medoids++ clustering. Because SIFT features represent shapes and edges, and because we use only the most abundant cluster medoids, this vector may represent the prevailing shapes and edges of nuclei in the image overall. This performs better than chance when used alone ($0.4102 \pm 0.0050$ vs $0.3967 \pm 0.0044$, $U = 100$, $p = 0.0001817$), and significantly improves performance when used in the context of 2,412 features, tissue covariate, and marker covariate ($0.6908 \pm 0.0021$ vs $0.6935 \pm 0.0029$, $U = 19.5$, $p = 0.02308$), though the effect is small.

**Combined distance function coefficient allows SIFT prioritization, but careful calibration not needed** Because SIFT feature vector magnitude is small, and the number of medoids we retain is small (i.e. 5), we ask if search would benefit from a SIFT coefficient greater than one in the combined distance function (i.e. $\alpha_{\text{SIFT}}$ in Eqn 4.14), to increase the relative contribution of SIFT features to search. We let $\alpha_{\text{SIFT}} = 5$. When combined with deep features, which we discuss later, we report a very small but significant improvement in search performance when $\alpha_{\text{SIFT}} = 5$ (0.6983±0.0016 vs 0.7006±0.0026, $U = 21.5$, $p = 0.03423$) but not when $\alpha_{\text{SIFT}} = 1$ (0.6983±0.0016 vs 0.6991±0.0023, $U = 34$, $p = 0.2408$). However, search performance is not significantly different depending on the choice of $\alpha_{\text{SIFT}}$ here (0.6991±0.0023 vs 0.7006±0.0026, $U = 30$, $p = 0.14$). Therefore, we do not observe compelling evidence in favor of selecting $\alpha_{\text{SIFT}}$ carefully, and caution that careful selection risks overfit. We note $\alpha_{\text{SIFT}} = 5$ may be viewed as a micro-optimization, but this effect is very small. It appears this effect is carried by $\text{SIFT}_{\text{k5}}$ rather than $\alpha_{\text{SIFT}}$, so we conclude it may be better to focus on feature engineering or learning, rather than coefficient selection. We report this to show $\alpha$ selection in the combined distance function has only a slight effect.

**Deep features synergize with other features, informing search more than nuclear SIFT features, but less than clinical covariates**

**For deep features, a combined distance function of L1 norm and Random Forest similarity also improve search** Because search performance improves when considering $\text{SIFT}_{\text{k5}}$ set features in a combined distance function, we ask if search performance improves when considering deep set features in a combined distance function (Eqn 4.14). We observe that naïvely concatenating the full 100-dimensional deep set representation with other features for Random Forest learning reduces performance (0.6533±0.0025 vs 0.5427±0.0025, $U = 100$, $p = 0.0001796$). We then select the interpretable top 3 most class-associated deep features (Figs 4.4, 4.5D1-3) and observe using these in a combined distance function improves search performance (0.6533±0.0025 vs 0.6602±0.0022, $U = 4$, $p = 0.0005773$). Performance remains improved when including the marker mention covariate (0.6908±0.0021 vs 0.6983±0.0016, $U = 0$, $p = 0.0001817$).

Given that both SIFT set features and deep set features both perform better in a combined distance function rather than within Random Forest similarity, we suspect our Random Forest parameters facilitate sensitive learning from few covariates, but may be oversensitive to a large number of set features that each take many different values. *Machine learning methods discussion* in the supplement discusses (Sec S4.5.11). We conclude an interpretable reduced representation of deep set features improves search performance when considered in a combined distance function, though this effect is small. We expect the effect to increase with more data, because deep learning can refine feature representations in a scalable data-driven manner, and because more advanced deep learning methods may be possible with more data.

**Deep features from supervised learning inform search more than nuclear SIFT features** Given that both $\text{SIFT}_{\text{k5}}$ and $\text{Deep}_3$ features improve search performance, we ask if pathology-specific $\text{Deep}_3$ features improve search performance more than pathology-agnostic $\text{SIFT}_{\text{k5}}$ features. Indeed, we observe $\text{Deep}_3$ features improve search performance significantly (i) more than SIFT features ($0.6983 \pm 0.0016$ vs $0.6376 \pm 0.0014$, $U = 100$, $p = 0.0001817$), (ii) more than $1 \times \text{SIFT}_{\text{k5}}$ features ($0.6983 \pm 0.0016$ vs $0.6935 \pm 0.0029$, $U = 94$, $p = 0.0003248$), and (iii) more than $5 \times \text{SIFT}_{\text{k5}}$ features ($0.6983 \pm 0.0016$ vs $0.6948 \pm 0.0032$, $U = 83.5$, $p = 0.01251$). We conclude learned pathology-specific deep features inform pathology search more than hand-engineered pathology agnostic SIFT features, though SIFT may cover nuclei.

**Deep features and SIFT features are complementary** To determine if SIFT and deep features represent non-overlapping concepts in pathology, we ask if

work, these near-duplicates should be automatically detected. Duplicates may artificially inflate performance metrics.

### Disease state annotations vary in text

Our dataset is only as good as the accuracy of the hashtags and diagnoses made by the collaborating pathologists and pathologists who comment on the cases. The more pathologists that contribute to the database, the higher the risk for errors and inconsistencies. Indeed we note some uses of the #bstpath hashtag to describe breast pathology (Section S4.5.6). We should remember the fun and voluntary nature of sharing cases on social media.

While our disease state text processing algorithms take a consensus vote among the pathologists discussing the patient case, these methods are not perfect, and our manual annotations to correct this may be incomplete. We hope that by sharing our data with the community, more corrections may be made, improving the quality of our dataset.

### Disease state evidence varies in images

For our Random Forest baselines, we crop images to convert rectangular images to be uniformly square for the machine learning. However, pathologists may include diagnostic information only at the extreme edges of an image that are cropped out. A case of this from B.X. involves a hydatid cyst in the extreme right of an image, which would be cropped out. This hydatid cyst indicates *Echinococ-*

---

Case at `https://twitter.com/BinXu16/status/980404471833313280` “Kudo to @drkennethtang @luishcruzc and @DrGeeONE The answer of this case can be seen in the right corner of the 3rd picture. Dx: Echinococcus (hydatid cyst) with necrotizing pneumonia, abscess, *cus* infection, so the case is nontumoral. Our set-based deep learning approach is an incomplete remedy to this problem, where we train on 224×224px image patches sampled throughout the original image, then test using 21 224×224px patches systematically sampled throughout the original image. Although this approach samples the entire image, the remedy is incomplete in that the ground truth is not uniform throughout the image. For example, it is only based on one corner of the image that there is histological evidence of nontumor disease. It may be helpful here to use more advanced methods, which make fewer assumptions about the ground truth and allow weaker supervision, but such methods may come at a cost of requiring more data than we have currently.

### Algorithm and labeling inaccuracies

We do not expect that our text classification algorithm (Fig S4.18) perfectly interprets disease state from the text associated with an image. Moreover, we also do not expect our manual annotation process to be perfect, e.g. some stains may be incorrectly labeled as H&E, IHC, etc. We manually curate 10,000+ images, so human error as low as 1% means a handful of images are incorrectly labeled, but there could be more. By sharing our data with more pathologists and data scientists, we intend to gather feedback and correct any inaccuracies here, then measure disease state classification performance changes.

---

and granulomatous inflammation. Additional high power pictures attached.”

### No automated quality control

Finally, the size of the dataset is both a blessing and a curse. A large and diverse dataset is required to provide the most benefit to computational pathology. However, quality control for such large datasets is most feasible if done automatically, and automated quality control cannot deal with all issues. For example, some pathology images include marks designating a particular pathologist as the contributor of that image. Other pathology images have been marked by pathologists with arrows and circles. Our automated quality control pipeline enables us to rapidly discissues. For example, some pathology images include marks designating a particular pathologist as the contributor of that image. Other pathology images have been marked by pathologists with arrows and circles. Our automated quality control pipeline enables us to rapidly discriminate pathology from non-pathology images, but is not able to address these other challenges. Future steps will need to be taken for more specialized quality control.

## S4.5.18 Supplementary Future directions

### Acquiring more data

The first step is to expand the size of this dataset by recruiting more pathologists via social media. With more data, we hope to improve performance on discriminations that were the most difficult (e.g. those involving gynecological pathology). More data may facilitate machine learning methods that discriminate between similar but less frequently used stains, such as H&E vs Diff-quik, rather than H&E vs IHC. More data might also enable us to distinguish particular organs or tissues within a histopathology tissue type, e.g. distinguish kidney tissue from bladder tissue. With increased sample size and increased tissue of origin granularity, it may be possible to predict metastatic tissue of origin. Finally, a larger dataset might also include more rare cases that can be useful for machine learning techniques that can support diagnoses.

### Expanded and specific hashtags

A second step is advocacy on social media, for (i) sharing normal tissue data, and (ii) expanded pathology hashtags. Normal tissue complements our existing "relatively unimportant" artifact and foreign body data, such as colloids and gauze, which are typically not prognostic of disease. Normal tissue also complements the description of tissue morphology in our data, if we tend to have only cancerous or diseased tissue. Separately, more descriptive hashtags may reduce our manual annotation burden, and obviate the need for us to ask pathologists to clarify what stain was used or what the tissue is. Moreover, molecular hashtags may complement the histology we see. However, we understand that for pathologists sharing cases on social media is probably a fun and voluntary activity, rather than a rigorous academic endeavor, so it may not be appropriate for us to suggest pathologists use terms from synoptic reporting in hashtag format in their Tweets. Moreover, the size of tweets is limited to 280 characters, so more than 3-4 hashtags per tweets is probably infeasible. Some pathologists are already close to this limit without using additional hashtags.

We encourage the adoption of hashtags that explicitly identify what stains or techniques are used (this is not an exhaustive list):

1. #he indicates there are one or more H&E-stained images in the tweets.
2. #ihc indicates there are one or more IHC-stained images in the tweets.
3. #pas indicates there are one or more periodic acid-Schiff images in the tweets.
4. #diffq indicates there are one or more diff-quik images in the tweets. There is a common misspelling of diff-quick, so our hashtag avoids this misspelling.
5. #gross indicates one or more gross section images are in the tweets. This is typically fresh cut tissue, e.g. an entire excised tumor or a large piece of an organ.
6. #macro indicates an unmagnified picture of a microscopy slide. Unfortunately, such pictures are occasionally referred to as gross.
7. #endo indicates one or more endoscopy images are in the tweets.
8. #ct indicates one or more CT scan images are in the tweets.
9. #xray indicates one or more X-ray images are in the tweets.

We encourage hashtags to describe not only the histological features of a case, but also the molecular features of a case. Again, this hashtag list is far from exhaustive.

1. #braf indicates the BRAF gene is known to be mutated, perhaps through sequencing.
2. #msi indicates micro-satellite instability, which again may be evident from sequencing.
3. #desmin indicates that the IHC used targets desmin.

We encourage the adoption of hashtags that explicitly identify any artifacts, art, or pathologist annotations/marks on the image.

1. #artifact or #artefact indicates there are artifacts or foreign bodies in one or more images, such as colloids, barium, sutures, Spongostan$^{\text{TM}}$, gauze, etc. We encourage the tweets message text to identify what the artifact or foreign body is.
2. #pathart is a hashtag in use today, but unfortunately it is used in two ways: (i) to identify naturally-occurring and unmodified pathology images that are "pretty" or "interesting" as natural works of art, and (ii) to identify images that have been modified by the pathologist herself/himself to be "funny" or "interesting". The trouble is (i) is "acceptable" pathology for analysis while (ii) is not. We advocate for the continued use of the #pathart hashtag, but with clarification, below:
3. #drawn or #annotated indicates the pathologist made hand-drawn marks on one or more images, such as arrows, circles, or artistic manipulations. Artistic manipulations may include drawing exclamation points, question marks, eyes, mouths, faces, skulls, cartoon bodies, etc on the image. So, "#pathart #drawn" is likely a pathology image with artistic drawn marks that prevents the image from being an "acceptable" pathology image for analysis, while "#pathart" without "#drawn" is likely a pathology image that is a naturally occurring unmodified histology image that is an "acceptable" pathology image for analysis.
4. Alternatively, "#pathfun" or "#pathdrawing" may refer artistically manipulated pathology pictures, leaving "#pathart" exclusively for naturally-occurring pathology that are "pretty" or "interesting" from an artistic per-

spective.

We encourage the adoption of hashtags that give other information about the image.

1. #pathbug is an existing hashtag that indicates a parasite or other co-occurring non-human organism is depicted in one or more images in the tweets. The #parasite tag is sometimes used instead.
2. #panel indicates one or more multi-panel images are in the tweets.

We encourage all adopted pathology-related hashtags to be registered in an ontology, e.g. `https://www.symplur.com/healthcare-hashtags/ontology/pathology/`. A hashtag ontology can standardize the hashtags used, which in turn can (i) help pathologists in the same subspecialty find each other, and (ii) simplify computational analyses of hashtag text.

# CHAPTER 5

# CONCLUSION

**Deep learning and whole slide images are staples in computational pathology** Computational pathology studies often apply deep learning to whole slide images, with applications including classification, detection, and segmentation[11]. Indeed, we similarly apply deep learning to whole slide images in chapter 2 and chapter 3. Whole slide images are high-quality micrographs, which are amenable for deep learning, and other automated analyses.

**Deep learning's performance strengths demand ample data** Deep learning excels at processing large amounts of raw data[151], which is both a strength and a weakness in computational pathology. Strong performance is possible without human-guided feature engineering, because the deep neural network can in a data-driven manner learn patterns from the raw data. A weakness is that strong performance often requires large amounts of diverse data, e.g. the ImageNet dataset has over 14 million images[51]. Datasets of this scale and diversity may not be readily available in most computational pathology studies.

**Data demands in light of pathology's multimodal and multi-tissue nature** To reduce the amount of data needed for deep learning to achieve acceptable performance, computational pathology studies often use pre-trained deep neural networks (i.e. "transfer learning"[3,4,5]), focus on a specific data modality (e.g. H&E, chest X-ray, etc), and focus on one specific tissue at a time (brain, breast, etc)[11]. Such focused study may not reflect a pathologist's multimodal approach to diagnosis in practice, which may consider together patient history, radiology, H&E, etc for a variety of tissues[21,24]. Restated, a pathologist may be expected to both consider multiple data modalities to render a single diagnosis, and for many different tissues accurately diagnose the full range of diseases from infection to malignancy. In contrast, often a single deep-learning-based classifier for computational pathology may only work with a single data modality and single tissue type[11]. Encouragingly however, there are recent developments in *multi-tissue* cancer-specific H&E-specific computational methods for whole slide images, where regions of interest are randomly-sampled[152] or pathologist-sampled[153]. As described in chapter 4, a recent method of ours is multi-tissue, pan-disease (i.e. all diseases from infection to malignancy), and multimodal (H&E with covariates for tissue type and marker stains e.g. IHC) for pathologist-sampled regions of interest[3].

**Crowd-sourcing makes data annotation and curation tractable** We applied a form of crowd-sourcing to address the challenge of obtaining a sufficiently large, diverse, and multimodal dataset for deep learning in computational pathology[3]. Specifically, we found the most success in respectfully asking pathologists worldwide to collaborate with us, according to protocols approved by an Institutional Review Board [IRB]. When we needed data, we asked pathologists. When we needed ideas, we asked pathologists. As more pathologists collaborated with us, our pathology data and pathologist interactions became more diverse and widely distributed. The hashtags and brevity encouraged by social media kept text annotations of photomicrographs simple, facilitating simple text analysis methods to determine ground truth labels for images. We avoided the "needle in a haystack problem", because pathologists would typically share regions of interest, rather than whole slide images. A caveat with our approach is that we manually curate each patient case from social media, which is a time-consuming. We shared this curation task with other students, to partially mitigate the time cost.

**Future directions** Whole slide images offer a complementary future direction, now that we have established methods for regions of interest[3]. Publicly available patient cohorts of whole slide images, like TCGA, offer whole exome sequencing data that are uncommon to find even summarized on social media. With some adaptation, we believe the global reach of our social media tools[3] may increase the impact of our mutation prediction tools[5]. Through social media, we may potentially add mutation prediction capabilities to hospitals worldwide where genomic sequencing is unavailable. Moreover, we found leveraging public data, e.g. H&E photomicrographs in published case reports available in PubMed Central, to qualitatively improve the utility of our patient case search tools[3]. Such published case reports offer a depth of analysis that is not possible to fully express in brief social media posts. We look forward to working with others to grow our nascent project, to connect pathologists worldwide and improve patient well-being, according to United Nations goals.

# BIBLIOGRAPHY

[1] United Nations. Good Health and Well-being, 2015.

[2] Transforming Our World: The 2030 Agenda for Sustainable Development. In W Rosa, editor, *A New Era in Global Health*. Springer Publishing Company, New York, NY, June 2017. ISBN 978-0-8261-9011-6 978-0-8261-9012-3. doi:10.1891/9780826190123.ap02.

[3] Schaumberg AJ, Juarez-Nicanor WC, Choudhury SJ, Pastrián LG, Pritt BS, Pozuelo MP, Sánchez RS, Ho K, Zahra N, Sener BD, Yip S, Xu B, Annavarapu SR, Morini A, Jones KA, Rosado-Orozco K, Mukhopadhyay S, Miguel C, Yang H, Rosen Y, Ali RH, Folaranmi OO, Gardner JM, Rusu C, Stayerman C, Gross J, Suleiman DE, Sirintrapun SJ, Aly M, and Fuchs TJ. Interpretable multimodal deep learning for real-time pan-tissue pan-disease pathology search on social media. *bioRxiv*, page 396663, March 2020. doi:10.1101/396663. Publisher: Cold Spring Harbor Laboratory Section: New Results.

[4] Schaumberg AJ, Joseph Sirintrapun S, Al-Ahmadie HA, Schüffler PJ, and Fuchs TJ. DeepScope: Nonintrusive Whole Slide Saliency Annotation and Prediction from Pathologists at the Microscope. In A Bracciali, G Caravagna, D Gilbert, and R Tagliaferri, editors, *Computational Intelligence Methods for Bioinformatics and Biostatistics*, Lecture Notes in Computer Science, pages 42–58. Springer International Publishing, Cham, 2017. ISBN 978-3-319-67834-4. doi:10.1007/978-3-319-67834-4_4.

[5] Schaumberg A, Rubin M, and Fuchs T. H&E-stained Whole Slide Deep Learning Predicts SPOP Mutation State in Prostate Cancer. *bioRxiv*, page 064279, July 2016. doi:10.1101/064279.

[6] Coudray N, Ocampo PS, Sakellaropoulos T, Narula N, Snuderl M, Fenyö D, Moreira A, Razavian N, and Tsirigos A. Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. *Nature medicine*, 24(10):1559–1567, October 2018. ISSN 1546-170X.

[7] Kather JN, Pearson AT, Halama N, Jager D, Krause J, Loosen SH, Marx A, Boor P, Tacke F, Neumann UP, Grabsch HI, Yoshikawa T, Brenner H, Chang-Claude J, Hoffmeister M, Trautwein C, and Luedde T. Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. *Nature Medicine*, 25(7):1054–1056, July 2019. ISSN 1546-170X. doi:10.1038/s41591-019-0462-y. Number: 7 Publisher: Nature Publishing Group.

[8] Arvaniti E, Fricker K, Moret M, Rupp N, Hermanns T, Fankhauser C, Wey N, Wild P, Rueschoff J, and Claassen M. Automated Gleason grading of prostate cancer tissue microarrays via deep learning. *bioRxiv*, January 2018. doi:10.1101/280024.

[9] Martin DR, Hanson JA, Gullapalli RR, Schultz FA, Sethi A, and Clark DP. A Deep Learning Convolutional Neural Network Can Recognize Common Patterns of Injury in Gastric Pathology. *Archives of Pathology & Laboratory Medicine*, 144(3):370–378, June 2019. ISSN 0003-9985. doi:10.5858/arpa.2019-0004-OA. Publisher: College of American Pathologists.

[10] Yi F, Yang L, Wang S, Guo L, Huang C, Xie Y, and Xiao G. Microvessel prediction in H&E Stained Pathology Images using fully convolutional

neural networks. *BMC Bioinformatics*, 19(1):64, February 2018. ISSN 1471-2105. doi:10.1186/s12859-018-2055-z.

[11] Litjens G, Kooi T, Bejnordi B, Setio A, Ciompi F, Ghafoorian M, van der Laak J, van Ginneken B, and SÃ¡nchez C. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60–88, 2017. ISSN 1361-8415. doi:10.1016/j.media.2017.07.005.

[12] Pell R, Oien K, Robinson M, Pitman H, Rajpoot N, Rittscher J, Snead D, and Verrill C. The use of digital pathology and image analysis in clinical trials. *The Journal of Pathology: Clinical Research*, 5(2):81–90, 2019. ISSN 2056-4538. doi:10.1002/cjp2.127. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/cjp2.127.

[13] Levine AB, Schlosser C, Grewal J, Coope R, Jones SJM, and Yip S. Rise of the Machines: Advances in Deep Learning for Cancer Diagnosis. *Trends in Cancer*, 5(3):157–169, March 2019. ISSN 2405-8033. doi:10.1016/j.trecan.2019.02.002.

[14] Acs B, Rantalainen M, and Hartman J. Artificial intelligence as the next step towards precision pathology. *Journal of Internal Medicine*, n/a(n/a), 2020. ISSN 1365-2796. doi:10.1111/joim.13030. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/joim.13030.

[15] Liangjun C, Honeine P, Hua Q, Jihong Z, and Xia S. Correntropy-based robust multilayer extreme learning machines. *Pattern Recognition*, 84:357–370, December 2018. ISSN 0031-3203. doi:10.1016/j.patcog.2018.07.011.

[16] PATHOLOGY VISIONS 2019. *Journal of Pathology Informatics*, 11, January 2020. ISSN 2229-5089.

[17] Müller H, Andrearczyk V, del Toro OJ, Dhrangadhariya A, Schaer R, and Atzori M. Studying Public Medical Images from the Open Access Literature and Social Networks for Model Training and Knowledge Extraction. In WH Cheng, J Kim, WT Chu, P Cui, JW Choi, MC Hu, and W De Neve, editors, *MultiMedia Modeling*, Lecture Notes in Computer Science, pages 553–564. Springer International Publishing, Cham, 2020. ISBN 978-3-030-37734-2. doi:10.1007/978-3-030-37734-2_45.

[18] Courtiol P, Maussion C, Moarii M, Pronier E, Pilcer S, Sefta M, Manceron P, Toldo S, Zaslavskiy M, Stang NL, Girard N, Elemento O, Nicholson AG, Blay JY, Galateau-Sallé F, Wainrib G, and Clozel T. Deep learning-based classification of mesothelioma improves prediction of patient outcome. *Nature Medicine*, 25(10):1519–1525, October 2019. ISSN 1546-170X. doi:10.1038/s41591-019-0583-3. Number: 10 Publisher: Nature Publishing Group.

[19] Li C, Chen H, Li X, Xu N, Hu Z, Xue D, Qi S, Ma H, Zhang L, and Sun H. A review for cervical histopathology image analysis using machine vision approaches. *Artificial Intelligence Review*, February 2020. ISSN 1573-7462. doi:10.1007/s10462-020-09808-7.

[20] Gamper J, Koohbanani NA, Benes K, Graham S, Jahanifar M, Khurram SA, Azam A, Hewitt K, and Rajpoot N. PanNuke Dataset Extension, Insights and Baselines. *arXiv:2003.10778 [cs, eess, q-bio]*, April 2020. ArXiv: 2003.10778.

[21] Hasleton P and Flieder DB. *Spencer's Pathology of the Lung*. Cambridge University Press, 2013.

[22] Tian S, Xiong Y, Liu H, Niu L, Guo J, Liao M, and Xiao SY. Pathological

study of the 2019 novel coronavirus disease (COVID-19) through post-mortem core biopsies. *Modern Pathology*, pages 1–8, April 2020. ISSN 1530-0285. doi:10.1038/s41379-020-0536-x. Publisher: Nature Publishing Group.

[23] Barton LM, Duval EJ, Stroberg E, Ghosh S, and Mukhopadhyay S. COVID-19 Autopsies, Oklahoma, USA. *American Journal of Clinical Pathology*, 2020. doi:10.1093/ajcp/aqaa062.

[24] Robinson BD, Mosquera JM, Ro JY, and Divatia M. *Precision Molecular Pathology of Prostate Cancer*. Springer, Cham, SWITZERLAND, 2018. ISBN 978-3-319-64096-9.

[25] Fuchs T and Buhmann J. Computational pathology: challenges and promises for tissue analysis. *Computerized medical imaging and graphics : the official journal of the Computerized Medical Imaging Society*, 35(7-8):515–530, October 2011. ISSN 1879-0771. doi:10.1016/j.compmedimag.2011.02.006.

[26] Brunye T, Carney P, Allison K, Shapiro L, Weaver D, and Elmore J. Eye Movements as an Index of Pathologist Visual Expertise: A Pilot Study. *PLoS ONE*, 9(8):e103447, August 2014. doi:10.1371/journal.pone.0103447.

[27] Raghunath V, Braxton M, Gagnon S, Brunye T, Allison K, Reisch L, Weaver D, Elmore J, and Shapiro L. Mouse cursor movement and eye tracking data as an indicator of pathologists' attention when viewing digital whole slide images. *Journal of pathology informatics*, 3, 2012. ISSN 2153-3539.

[28] Romo D, Romero E, and Gonzalez F. Learning regions of interest from low level maps in virtual microscopy. *Diagnostic Pathology*, 6(Suppl 1):S22, 2011. ISSN 1746-1596. doi:10.1186/1746-1596-6-s1-s22.

[29] Mercan E, Aksoy S, Shapiro L, Weaver D, Brunye T, and Elmore J. Localization of Diagnostically Relevant Regions of Interest in Whole Slide Images. pages 1179–1184. IEEE, August 2014. ISBN 1051-4651. doi: 10.1109/icpr.2014.212.

[30] Krupinski E, Tillack A, Richter L, Henderson J, Bhattacharyya A, Scott K, Graham A, Descour M, Davis J, and Weinstein R. Eye-movement study and human performance using telepathology virtual slides. Implications for medical education and differences with experience. *Human Pathology*, 37(12):1543–1556, December 2006. ISSN 00468177. doi: 10.1016/j.humpath.2006.08.024.

[31] Ball R and North C. The effects of peripheral vision and physical navigation on large scale visualization. *Proceedings of Graphics Interface 2008*, pages 9–16, 2008. ISSN 978-1-56881-423-0.

[32] Ball R, North C, and Bowman D. Move to Improve: Promoting Physical Navigation to Increase User Performance with Large Displays. pages 191–200. ACM, 2007. ISBN 978-1-59593-593-9. doi:10.1145/1240624.1240656.

[33] Randell R, Ambepitiya T, Mello-Thoms C, Ruddle R, Brettle D, Thomas R, and Treanor D. Effect of Display Resolution on Time to Diagnosis with Virtual Pathology Slides in a Systematic Search Task. *Journal of Digital Imaging*, 28(1):68–76, 2015. doi:10.1007/s10278-014-9726-8.

[34] Shupp L, Ball R, Yost B, Booker J, and North C. Evaluation of viewport size and curvature of large, high-resolution displays. pages 123–130. Canadian Information Processing Society, 2006. ISBN 1-56881-308-2.

[35] Starkweather G. 58.4: DSHARP — A Wide Screen Multi-Projector Dis-

play. *SID Symposium Digest of Technical Papers*, 34(1):1535–1537, May 2003. doi:10.1889/1.1832577.

[36] Parwani A, Hassell L, Glassy E, and Pantanowitz L. Regulatory barriers surrounding the use of whole slide imaging in the United States of America. *Journal of pathology informatics*, 5(1), 2014. ISSN 2229-5089. doi: 10.4103/2153-3539.143325.

[37] Begelman G, Lifshits M, and Rivlin E. Visual positioning of previously defined ROIs on microscopic slides. *IEEE Transactions on Information Technology in Biomedicine*, 10(1):42–50, January 2006. ISSN 1089-7771. doi: 10.1109/titb.2005.856856.

[38] Just M and Carpenter P. A theory of reading: From eye fixations to comprehension. *Psychological Review*, 87(4):329–354, 1980. ISSN 0033-295X. doi:10.1037/0033-295x.87.4.329.

[39] Erwin D. *The Interface of Language, Vision, and Action*. The Interface of Language, Vision, and Action. Routledge, June 2004. ISBN 978-0-203-48843-0. doi:10.4324/9780203488430.

[40] Eivazi S, Bednarik R, Leinonen V, Mikael von und zu Fraunberg, and Jaaskelainen J. Embedding an Eye Tracker Into a Surgical Microscope: Requirements, Design, and Implementation. *IEEE Sensors Journal*, 16(7):2070–2078, April 2016. ISSN 1530-437X. doi:10.1109/jsen.2015.2501237.

[41] Keerativittayanun S, Rakjaeng K, Kondo T, Kongprawechnon W, Tungpimolrut K, and Leelasawassuk T. Eye tracking system for Ophthalmic

Operating Microscope. pages 653–656. IEEE, August 2009. ISBN 978-4-907764-34-0.

[42] Goode A, Gilbert B, Harkes J, Jukic D, and Satyanarayanan M. OpenSlide: A vendor-neutral software foundation for digital pathology. *Journal of pathology informatics*, 4, 2013. ISSN 2229-5089. doi:10.4103/2153-3539.119005.

[43] Farneback G. Two-frame Motion Estimation Based on Polynomial Expansion. pages 363–370. Springer-Verlag, 2003. ISBN 3-540-40601-8.

[44] Farneback G. *Polynomial Expansion for Orientation and Motion Estimation*. Ph.D. thesis, Linkoping University, Sweden, 2002.

[45] Schneider C, Rasband W, and Eliceiri K. NIH Image to ImageJ: 25 years of image analysis. *Nature methods*, 9(7):671–675, July 2012. ISSN 1548-7105.

[46] Bay H, Tuytelaars T, and Van Gool L. SURF: Speeded Up Robust Features. In A Leonardis, H Bischof, and A Pinz, editors, *Computer Vision – ECCV 2006*, volume 3951, pages 404–417. Springer Berlin Heidelberg, 2006. ISBN 978-3-540-33832-1. doi:10.1007/11744023_32.

[47] Fischler M and Bolles R. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, June 1981.

[48] Srivastava N, Hinton G, Krizhevsky A, Sutskever I, and Salakhutdinov R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. volume 15, pages 1929–1958. June 2014.

[49] Jia Y, Shelhamer E, Donahue J, Karayev S, Long J, Girshick R, Guadarrama S, and Darrell T. Caffe: Convolutional Architecture for Fast Feature Embedding. June 2014.

[50] Krizhevsky A, Sutskever I, and Hinton G. Imagenet classification with deep convolutional neural networks. 2012.

[51] Deng J, Dong W, Socher R, Li LJ, Li K, and Fei-Fei L. ImageNet: A large-scale hierarchical image database. pages 248–255. IEEE, June 2009. ISBN 978-1-4244-3992-8. doi:10.1109/cvpr.2009.5206848.

[52] Thorvaldsdóttir H, Robinson J, and Mesirov J. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in bioinformatics*, 14(2):178–192, March 2013. ISSN 1477-4054. doi:10.1093/bib/bbs017.

[53] Robinson J, Thorvaldsdóttir H, Winckler W, Guttman M, Lander E, Getz G, and Mesirov J. Integrative genomics viewer. *Nature biotechnology*, 29(1):24–26, January 2011. ISSN 1546-1696. doi:10.1038/nbt.1754.

[54] Mosquera JM, Perner S, Demichelis F, Kim R, Hofer M, Mertz K, Paris P, Simko J, Collins C, Bismar T, Chinnaiyan A, and Rubin M. Morphological features of TMPRSS2-ERG gene fusion prostate cancer. *The Journal of pathology*, 212(1):91–101, May 2007. ISSN 0022-3417. doi:10.1002/path.2154.

[55] Cheng D, Mitchell T, Zehir A, Shah R, Benayed R, Syed A, Chandramohan R, Liu ZY, Won H, Scott S, Brannon R, O'Reilly C, Sadowska J, Casanova J, Yannes A, Hechtman J, Yao J, Song W, Ross D, Oultache A, Dogan S, Borsu L, Hameed M, Nafa K, Arcila M, Ladanyi M, and Berger M. Memorial

Sloan Kettering-Integrated Mutation Profiling of Actionable Cancer Targets (MSK-IMPACT): A Hybridization Capture-Based Next-Generation Sequencing Clinical Assay for Solid Tumor Molecular Oncology. *The Journal of molecular diagnostics : JMD*, 17(3):251–264, May 2015. ISSN 1943-7811.

[56] Barbieri C, Baca S, Lawrence M, Demichelis F, Blattner M, Theurillat JP, White T, Stojanov P, Van Allen E, Stransky N, Nickerson E, Chae SS, Boysen G, Auclair D, Onofrio R, Park K, Kitabayashi N, MacDonald T, Sheikh K, Vuong T, Guiducci C, Cibulskis K, Sivachenko A, Carter S, Saksena G, Voet D, Hussain W, Ramos A, Winckler W, Redman M, et al. Exome sequencing identifies recurrent SPOP, FOXA1 and MED12 mutations in prostate cancer. *Nature Genetics*, 44(6):685–689, May 2012. ISSN 1061-4036. doi:10.1038/ng.2279.

[57] Hakimi A, Tickoo S, Jacobsen A, Sarungbam J, Sfakianos J, Sato Y, Morikawa T, Kume H, Fukayama M, Homma Y, Chen YB, Sankin A, Mano R, Coleman J, Russo P, Ogawa S, Sander C, Hsieh J, and Reuter V. TCEB1-mutated renal cell carcinoma: a distinct genomic and morphological subtype. *Modern pathology : an official journal of the United States and Canadian Academy of Pathology, Inc*, 28(6):845–853, June 2015. ISSN 1530-0285.

[58] Weisman P, Ng C, Brogi E, Eisenberg R, Won H, Piscuoglio S, De Filippo M, Ioris R, Akram M, Norton L, Weigelt B, Berger M, Reis-Filho J, and Wen H. Genetic alterations of triple negative breast cancer by targeted next-generation sequencing and correlation with tumor morphology. *Mod Pathol*, 29(5):476–488, May 2016. ISSN 0893-3952. doi:10.1038/modpathol.2016.39.

[59] Budinska E, Bosman F, and Popovici V. Experiments in molecular subtype recognition based on histopathology images. pages 1168–1172. IEEE, April 2016. doi:10.1109/%20isbi.2016.7493474.

[60] Zhong Q, Rüschoff J, Guo T, Gabrani M, Schüffler P, Rechsteiner M, Liu Y, Fuchs T, Rupp N, Fankhauser C, Buhmann J, Perner S, Poyet C, Blattner M, Soldini D, Moch H, Rubin M, Noske A, Rüschoff J, Haffner M, Jochum W, and Wild P. Image-based computational quantification and visualization of genetic alterations and tumour heterogeneity. *Scientific Reports*, 6:24146, April 2016. ISSN 2045-2322. doi:10.1038/srep24146.

[61] Natrajan R, Sailem H, Mardakheh F, Arias Garcia M, Tape C, Dowsett M, Bakal C, and Yuan Y. Microenvironmental Heterogeneity Parallels Breast Cancer Progression: A Histology–Genomic Integration Analysis. *PLoS Med*, 13(2):e1001961, February 2016. doi:10.1371/journal.pmed.1001961.

[62] He K, Zhang X, Ren S, and Sun J. Deep Residual Learning for Image Recognition. December 2015.

[63] Qi X, Wang D, Rodero I, Diaz-Montes J, Gensure R, Xing F, Zhong H, Goodell L, Parashar M, Foran D, and Yang L. Content-based histopathology image retrieval using CometCloud. *BMC Bioinformatics*, 15(1):287, August 2014. ISSN 1471-2105. doi:10.1186/1471-2105-15-287.

[64] Giesen C, Wang H, Schapiro D, Zivanovic N, Jacobs A, Hattendorf B, Schüffler P, Grolimund D, Buhmann J, Brandt S, Varga Z, Wild P, Günther D, and Bodenmiller B. Highly multiplexed imaging of tumor tissues with subcellular resolution by mass cytometry. *Nature methods*, 11(4):417–422, April 2014. ISSN 1548-7105.

[65] An J, Wang C, Deng Y, Yu L, and Huang H. Destruction of full-length androgen receptor by wild-type SPOP, but not prostate-cancer-associated mutants. *Cell reports*, 6(4):657–669, February 2014. ISSN 2211-1247.

[66] Urbinati G, de Waziers I, Slamiç M, Foussignière T, Ali H, Desmaële D, Couvreur P, and Massaad-Massade L. Knocking Down TMPRSS2-ERG Fusion Oncogene by siRNA Could be an Alternative Treatment to Flutamide. *Molecular therapy. Nucleic acids*, 5, 2016. ISSN 2162-2531.

[67] Schaumberg A, Sirintrapun S, Al-Ahmadie H, Schueffler P, and Fuchs T. DeepScope: Nonintrusive Whole Slide Saliency Annotation and Prediction from Pathologists at the Microscope. *bioRxiv*, page 097246, December 2016. doi:10.1101/097246.

[68] Krogh A and Vedelsby J. Neural network ensembles, cross validation, and active learning. *Advances in neural information processing systems*, 7:231–238, 1995.

[69] Sollich P and Krogh A. Learning with ensembles: How over-fitting can be useful. volume 8, pages 190–196. 1996.

[70] Breiman L. Random Forests. *Machine Learning*, 45(1):5–32, 2001. doi:10.1023/a%253a1010933404324.

[71] Mosteller F and Fisher R. Questions and Answers. *The American Statistician*, 2:30–31, October 1948.

[72] Poole W, Gibbs D, Shmulevich I, Bernard B, and Knijnenburg T. Combining dependent P-values with an empirical adaptation of Brown's method. *Bioinformatics (Oxford, England)*, 32(17), September 2016. ISSN 1367-4811.

[73] Kost J and McDermott M. Combining dependent P-values. *Statistics & Probability Letters*, 60(2):183–190, November 2002. ISSN 01677152. doi: 10.1016/s0167-7152(02)00310-3.

[74] Cerami E, Gao J, Dogrusoz U, Gross B, Sumer S, Aksoy B, Jacobsen A, Byrne C, Heuer M, Larsson E, Antipin Y, Reva B, Goldberg A, Sander C, and Schultz N. The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data. *Cancer Discovery*, 2(5):401–404, May 2012. ISSN 2159-8290. doi:10.1158/2159-8290.cd-12-0095.

[75] Schueffler P, Fuchs T, Cheng Soon, Wild P, Rupp N, and Buhmann J. TMARKER: A free software toolkit for histopathological cell counting and staining estimation. *Journal of pathology informatics*, 4(Suppl), 2013. ISSN 2229-5089.

[76] Nix J, Gardner J, Costa F, Soares A, Rodriguez F, Moore B, Martinez-Lage M, Ahlawat S, Gokden M, and Anthony D. Neuropathology Education Using Social Media. *Journal of neuropathology and experimental neurology*, 77(6):454–460, June 2018. ISSN 1554-6578.

[77] Crane G and Gardner J. Pathology Image-Sharing on Social Media: Recommendations for Protecting Privacy While Motivating Education. *AMA journal of ethics*, 18(8):817–825, August 2016. ISSN 2376-6980.

[78] Dirilenoglu F and Önal B. A welcoming guide to social media for cytopathologists: Tips, tricks, and the best practices of social cytopathology. *CytoJournal*, 16:4–4, February 2019. ISSN 1742-6413. doi:10.4103/cytojournal.cytojournal_1_18.

[79] Gardner J and Allen T. Keep Calm and Tweet On: Legal and Ethical Considerations for Pathologists Using Social Media. *Archives of pathology & laboratory medicine*, 143:75–80, August 2018. ISSN 1543-2165.

[80] Lepe M, Wu R, Oltulu P, Alex D, Canepa M, Deeken A, Doxtader E, Fitzhugh V, Gibier JB, Jain D, Janaki N, Jelinek A, Kumar S, Labiano T, L'Imperio V, Michael C, Mukhopadhyay S, Pagni F, Panizo A, Pijuan L, Quintana L, Roy-Chowdhuri S, Sanchez-Font A, Valero I, Sauter J, Skipper D, Spruill L, Torous V, Gardner J, and Jiang X. #EBUSTwitter: Novel Use of Social Media for Conception, Coordination and Completion of an International, Multi-Center Pathology Study. *Journal of the American Society of Cytopathology*, 7(5):S88–S89, September 2018. ISSN 2213-2945. doi:10.1016/j.jasc.2018.06.015.

[81] Liu Y, Chen PHC, Krause J, and Peng L. How to Read Articles That Use Machine Learning: Users' Guides to the Medical Literature. *JAMA*, 322(18):1806–1816, November 2019. ISSN 0098-7484. doi:10.1001/jama.2019.16489.

[82] Montalto MC. An industry perspective: An update on the adoption of whole slide imaging. *Journal of Pathology Informatics*, 7, April 2016. ISSN 2229-5089. doi:10.4103/2153-3539.180014.

[83] Zarella MD, Bowman; D, Aeffner F, Farahani N, Xthona; A, Absar SF, Parwani A, Bui M, and Hartman DJ. A Practical Guide to Whole Slide Imaging: A White Paper From the Digital Pathology Association. *Archives of Pathology & Laboratory Medicine*, 143(2):222–234, October 2018. ISSN 0003-9985. doi:10.5858/arpa.2018-0343-RA.

[84] Krizhevsky A. *Learning Multiple Layers of Features from Tiny Images*. Mas-

ter's thesis, University of Toronto, Department of Computer Science, April 2009.

[85] Christiansen E, Yang S, Ando M, Javaherian A, Skibinski G, Lipnick S, Mount E, O'Neil A, Shah K, Lee A, Goyal P, Fedus W, Poplin R, Esteva A, Berndl M, Rubin L, Nelson P, and Finkbeiner S. In Silico Labeling: Predicting Fluorescent Labels in Unlabeled Images. *Cell*, 173(3), April 2018. ISSN 1097-4172.

[86] Bayramoglu N and Heikkila J. Transfer Learning for Cell Nuclei Classification in Histopathology Images. pages 532–539. Springer International Publishing, 2016. ISBN 978-3-319-49409-8. doi:10.1007/978-3-319-49409-8_46.

[87] Tajbakhsh N, Shin J, Gurudu S, Hurst T, Kendall C, Gotway M, and Jianming Liang. Convolutional Neural Networks for Medical Image Analysis: Full Training or Fine Tuning? *IEEE transactions on medical imaging*, 35(5):1299–1312, May 2016. ISSN 1558-254X.

[88] Komura D, Fukuta K, Tominaga K, Kawabe A, Koda H, Suzuki R, Konishi H, Umezaki T, Harada T, and Ishikawa S. Luigi: Large-scale histopathological image retrieval system using deep texture representations. *bioRxiv*, page 345785, July 2018. doi:10.1101/345785.

[89] Hegde N, Hipp JD, Liu Y, Emmert-Buck M, Reif E, Smilkov D, Terry M, Cai CJ, Amin MB, Mermel CH, Nelson PQ, Peng LH, Corrado GS, and Stumpe MC. Similar image search for histopathology: SMILY. *npj Digital Medicine*, 2(1):1–9, June 2019. ISSN 2398-6352. doi:10.1038/s41746-019-0131-z.

[90] Otalora S, Schaer R, Atzori M, Jimenez del Toro O, and Muller H. Deep learning based retrieval system for gigapixel histopathology cases and open access literature. *bioRxiv*, January 2018. doi:10.1101/408237.

[91] Veta M, Heng Y, Stathonikos N, Bejnordi B, Beca F, Wollmann T, Rohr K, Shah M, Wang D, Rousson M, Hedlund M, Tellez D, Ciompi F, Zerhouni E, Lanyi D, Viana M, Kovalev V, Liauchuk V, Phoulady H, Qaiser T, Graham S, Rajpoot N, Sjöblom E, Molin J, Paeng K, Hwang S, Park S, Jia Z, Chang E, Xu Y, et al. Predicting breast tumor proliferation from whole-slide images: the TUPAC16 challenge. July 2018. doi:10.1016/j.media.2019.02.012.

[92] Jang SJ, Gardner JM, and Ro JY. Diagnostic approach and prognostic factors of cancers. *Advances in Anatomic Pathology*, 18(2):165–172, March 2011. ISSN 1533-4031. doi:10.1097/PAP.0b013e31820cb39e.

[93] Rajkomar A, Dean J, and Kohane I. Machine Learning in Medicine. *New England Journal of Medicine*, 380(14):1347–1358, April 2019. ISSN 0028-4793. doi:10.1056/NEJMra1814259.

[94] Begoli E, Bhattacharya T, and Kusnezov D. The need for uncertainty quantification in machine-assisted medical decision making. *Nature Machine Intelligence*, 1(1):20–23, January 2019. ISSN 2522-5839. doi:10.1038/s42256-018-0004-1.

[95] Marshall A and Davies A. Uber's Self-Driving Car Didn't Know Pedestrians Could Jaywalk. *Wired*, November 2019. ISSN 1059-1028.

[96] Vlasic B and Boudette NE. Self-Driving Tesla Was Involved in Fatal Crash, U.S. Says. *The New York Times*, June 2016. ISSN 0362-4331.

[97] Zhang H, Cisse M, Dauphin YN, and Lopez-Paz D. mixup: Beyond Empirical Risk Minimization. *arXiv:1710.09412 [cs, stat]*, April 2018. ArXiv: 1710.09412.

[98] Breiman L, Friedman J, Olshen R, and Stone C. *Classification and regression trees*. Chapman and Hall, New York, 1984. ISBN 1-351-46049-8.

[99] Cabitza F, Rasoini R, and Gensini GF. Unintended Consequences of Machine Learning in Medicine. *JAMA*, 318(6):517–518, August 2017. ISSN 0098-7484. doi:10.1001/jama.2017.7797.

[100] Caruana R, Lou Y, Gehrke J, Koch P, Sturm M, and Elhadad N. Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, pages 1721–1730. ACM, New York, NY, USA, 2015. ISBN 978-1-4503-3664-2. doi: 10.1145/2783258.2788613. Event-place: Sydney, NSW, Australia.

[101] McInnes L, Healy J, and Melville J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. February 2018.

[102] Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, and Witten I. The WEKA Data Mining Software: An Update. *SIGKDD Explor. Newsl.*, 11(1):10–18, November 2009. doi:10.1145/1656274.1656278.

[103] Chollet F and others. Keras, 2015. Https://keras.io.

[104] Liaw A and Wiener M. Classification and Regression by randomForest. *R News*, 2(3):18–22, 2002. Https://cran.r-project.org/web/packages/randomForest/index.html.

[105] Konopka T. umap: Uniform Manifold Approximation and Projection, January 2020.

[106] Doshi-Velez F and Perlis RH. Evaluating Machine Learning Articles. *JAMA*, 322(18):1777–1779, November 2019. ISSN 0098-7484. doi:10.1001/jama.2019.17304.

[107] Esteva A, Kuprel B, Novoa R, Ko J, Swetter S, Blau H, and Thrun S. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118, January 2017. ISSN 0028-0836. doi:10.1038/nature21056.

[108] De Fauw J, Ledsam J, Romera-Paredes B, Nikolov S, Tomasev N, Blackwell S, Askham H, Glorot X, Oâ€™Donoghue B, Visentin D, van den Driessche G, Lakshminarayanan B, Meyer C, Mackinder F, Bouton S, Ayoub K, Chopra R, King D, Karthikesalingam A, Hughes Ca, Raine R, Hughes J, Sim D, Egan C, Tufail A, Montgomery H, Hassabis D, Rees G, Back T, Khaw P, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature Medicine*, 2018. ISSN 1546-170X. doi:10.1038/s41591-018-0107-6.

[109] Campanella G, Hanna MG, Geneslaw L, Miraflor A, Werneck Krauss Silva V, Busam KJ, Brogi E, Reuter VE, Klimstra DS, and Fuchs TJ. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature Medicine*, 25(8):1301–1309, 2019. ISSN 1546-170X. doi:10.1038/s41591-019-0508-1.

[110] Mehta N, Alomari RS, and Chaudhary V. Content based sub-image retrieval system for high resolution pathology images using salient interest points. *Conference proceedings: ... Annual International Conference of the IEEE*

*Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual Conference*, 2009:3719–3722, 2009. ISSN 1557-170X. doi:10.1109/IEMBS.2009.5334811.

[111] Zheng Y, Jiang Z, Zhang H, Xie F, Ma Y, Shi H, and Zhao Y. Histopathological Whole Slide Image Analysis Using Context-Based CBIR. *IEEE Transactions on Medical Imaging*, 37(7):1641–1652, July 2018. ISSN 0278-0062, 1558-254X. doi:10.1109/TMI.2018.2796130.

[112] Lowe D. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004. doi:10.1023/b%253avisi.0000029664.99615.94.

[113] Nazeran H, Rice F, Moran W, and Skinner J. Biomedical image processing in pathology: a review. *Australasian Physical & Engineering Sciences in Medicine*, 18(1):26–38, March 1995. ISSN 0158-9938.

[114] Linder N, Konsti J, Turkki R, Rahtu E, Lundin M, Nordling S, Haglund C, Ahonen T, Pietikäinen M, and Lundin J. Identification of tumor epithelium and stroma in tissue microarrays using texture analysis. *Diagnostic Pathology*, 7(1):22, March 2012. ISSN 1746-1596. doi:10.1186/1746-1596-7-22.

[115] Kather JN, Weis CA, Bianconi F, Melchers SM, Schad LR, Gaiser T, Marx A, and Zöllner FG. Multi-class texture analysis in colorectal cancer histology. *Scientific Reports*, 6:27988, June 2016. ISSN 2045-2322. doi:10.1038/srep27988.

[116] Tefik T, Sanli O, Oktar T, Yucel OB, Ozluk Y, and Kilicaslan I. Oxidized regenerated cellulose granuloma mimicking recurrent mass lesion after

laparoscopic nephron sparing surgery. *International journal of surgery case reports*, 3(6):227–230, 2012. ISSN 2210-2612.

[117] Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, and Rabinovich A. Going Deeper with Convolutions. September 2014.

[118] Oltulu P, Mannan AASR, and Gardner J. Effective use of Twitter and Facebook in pathology practice. *Human pathology*, 73:128–143, March 2018. ISSN 1532-8392.

[119] Lux M and Chatzichristofis S. Lire: Lucene Image Retrieval: An Extensible Java CBIR Library. pages 1085–1088. ACM, 2008. ISBN 978-1-60558-303-7. doi:10.1145/1459359.1459577.

[120] Chatzichristofis S and Boutalis Y. CEDD: Color and Edge Directivity Descriptor: A Compact Descriptor for Image Indexing and Retrieval. In A Gasteratos, M Vincze, and J Tsotsos, editors, *Computer Vision Systems*, volume 5008, pages 312–322. Springer Berlin Heidelberg, 2008. doi:10.1007/978-3-540-79547-6_30.

[121] Huang J, Kumar S, Mitra M, Zhu WJ, and Zabih R. Image indexing using color correlograms. In *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, volume 0, pages 762–768. IEEE, June 1997. ISBN 0-8186-7822-4. doi:10.1109/cvpr.1997.609412.

[122] Chatzichristofis S and Boutalis Y. FCTH: Fuzzy Color and Texture Histogram - A Low Level Feature for Accurate Image Retrieval. In *Image Analysis for Multimedia Interactive Services, 2008. WIAMIS '08. Ninth Inter-*

*national Workshop on*, pages 191–196. IEEE, May 2008. ISBN 978-0-7695-3344-5. doi:10.1109/wiamis.2008.24.

[123] Ojala T, Pietikainen M, and Harwood D. Performance evaluation of texture measures with classification based on Kullback discrimination of distributions. In *Proceedings of 12th International Conference on Pattern Recognition*, volume 1, pages 582–585 vol.1. October 1994. doi:10.1109/ICPR.1994.576366.

[124] Ahonen T, Hadid A, and Pietikainen M. Face Description with Local Binary Patterns: Application to Face Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(12):2037–2041, December 2006. ISSN 1939-3539. doi:10.1109/TPAMI.2006.244.

[125] Bosch A, Zisserman A, and Munoz X. Representing Shape with a Spatial Pyramid Kernel. In *Proceedings of the 6th ACM International Conference on Image and Video Retrieval*, CIVR '07, pages 401–408. Association for Computing Machinery, New York, NY, USA, 2007. ISBN 978-1-59593-733-9. doi:10.1145/1282280.1282340. Event-place: Amsterdam, The Netherlands.

[126] Ojala T, Pietikainen M, and Maenpaa T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7):971–987, July 2002. ISSN 0162-8828. doi:10.1109/TPAMI.2002.1017623.

[127] Tamura H, Mori S, and Yamawaki T. Textural Features Corresponding to Visual Perception. *IEEE Transactions on Systems, Man, and Cybernetics*, 8(6):460–473, June 1978. ISSN 0018-9472. doi:10.1109/TSMC.1978.4309999.

[128] Zaheer M, Kottur S, Ravanbakhsh S, Poczos B, Salakhutdinov RR, and Smola AJ. Deep Sets. In I Guyon, UV Luxburg, S Bengio, H Wallach, R Fergus, S Vishwanathan, and R Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 3391–3401. Curran Associates, Inc., 2017. Https://arxiv.org/abs/1703.06114.

[129] Mikolov T, Chen K, Corrado G, and Dean J. Efficient Estimation of Word Representations in Vector Space. *arXiv:1301.3781 [cs]*, September 2013. ArXiv: 1301.3781.

[130] Kim H and Loh WY. Classification Trees With Unbiased Multiway Splits. *Journal of the American Statistical Association*, 96(454):589–604, June 2001. ISSN 0162-1459. doi:10.1198/016214501753168271.

[131] Maaten Lvd and Hinton G. Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9(Nov):2579–2605, 2008. ISSN ISSN 1533-7928.

[132] Bai S. Growing random forest on deep convolutional neural networks for scene categorization. *Expert Systems with Applications*, 71:279–287, April 2017. ISSN 0957-4174. doi:10.1016/j.eswa.2016.10.038.

[133] Sutskever I, Martens J, and Dahl G. On the importance of initialization and momentum in deep learning. page 9, 2013.

[134] Nesterov Y. A method of solving a convex programming problem with convergence rate O(1/sqr(k)). In *Soviet Mathematics Doklady*, volume 27, pages 372–376. 1983.

[135] Nesterov Y. *Introductory Lectures on Convex Optimization: A Basic Course*. Applied Optimization. Springer US, 2004. ISBN 978-1-4020-7553-7. doi: 10.1007/978-1-4419-8853-9.

[136] Klimek MD and Perelstein M. Neural Network-Based Approach to Phase Space Integration. *arXiv:1810.11509 [hep-ex, physics:hep-ph, physics:physics, stat]*, October 2018. ArXiv: 1810.11509.

[137] Hochreiter S and Schmidhuber J. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780, November 1997. ISSN 0899-7667. doi: 10.1162/neco.1997.9.8.1735.

[138] Hochreiter S, Bengio Y, Frasconi P, and Schmidhuber J. *Gradient flow in recurrent nets: the difficulty of learning long-term dependencies*. IEEE Press, 2001.

[139] Lin K, Yang HF, Hsiao JH, and Chen CS. Deep learning of binary hash codes for fast image retrieval. pages 27–35. IEEE, June 2015. doi:10.1109/cvprw.2015.7301269.

[140] Sivic J and Zisserman A. Efficient Visual Search of Videos Cast as Text Retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(4):591–606, April 2009. ISSN 0162-8828, 2160-9292, 1939-3539. doi:10.1109/TPAMI.2008.111.

[141] Zhang X, Liu W, Dundar M, Badve S, and Zhang S. Towards Large-Scale Histopathological Image Analysis: Hashing-Based Image Retrieval. *IEEE Transactions on Medical Imaging*, 34(2):496–506, February 2015. ISSN 0278-0062, 1558-254X. doi:10.1109/TMI.2014.2361481.

[142] Wang JZ. Pathfinder: multiresolution region-based searching of pathology images using IRM. *Proceedings of the AMIA Symposium*, pages 883–887, 2000. ISSN 1531-605X.

[143] Mosquera-Lopez C, Agaian S, Velez-Hoyos A, and Thompson I. Computer-Aided Prostate Cancer Diagnosis From Digitized Histopathology: A Review on Texture-Based Systems. *IEEE Reviews in Biomedical Engineering*, 8:98–113, 2015. ISSN 1937-3333, 1941-1189. doi:10.1109/RBME.2014.2340401.

[144] Sridhar A, Doyle S, and Madabhushi A. Content-based image retrieval of digitized histopathology in boosted spectrally embedded spaces. *Journal of pathology informatics*, 6, 2015. ISSN 2229-5089. doi:10.4103/2153-3539.159441.

[145] Gleason Donald F, Mellinger George T, and Group TVACUR. Prediction of Prognosis for Prostatic Adenocarcinoma by Combined Histological Grading and Clinical Staging. *Journal of Urology*, 111(1):58–64, January 1974. doi:10.1016/S0022-5347(17)59889-4.

[146] Markonis D, Schaer R, de Herrera AGS, and Müller H. The Parallel Distributed Image Search Engine (ParaDISE). *arXiv:1701.05596 [cs]*, January 2017. ArXiv: 1701.05596.

[147] Peng T, Boxberg M, Weichert W, Navab N, and Marr C. Multi-task Learning of a Deep K-Nearest Neighbour Network for Histopathological Image Classification and Retrieval. In D Shen, T Liu, TM Peters, LH Staib, C Essert, S Zhou, PT Yap, and A Khan, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, Lecture Notes in Computer Science, pages 676–684. Springer International Publishing, Cham, 2019. ISBN 978-3-030-32239-7. doi:10.1007/978-3-030-32239-7_75.

[148] Cao Z, Long M, Wang J, and Yu PS. HashNet: Deep Learning to Hash by Continuation. *arXiv:1702.00758 [cs]*, July 2017. ArXiv: 1702.00758.

[149] Cao Y, Long M, Liu B, and Wang J. Deep Cauchy Hashing for Hamming Space Retrieval. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1229–1237. IEEE, Salt Lake City, UT, June 2018. ISBN 978-1-5386-6420-9. doi:10.1109/CVPR.2018.00134.

[150] Commandeur F, Slomka PJ, Goeller M, Chen X, Cadet S, Razipour A, McElhinney P, Gransar H, Cantu S, Miller RJH, Rozanski A, Achenbach S, Tamarappoo BK, Berman DS, and Dey D. Machine learning to predict the long-term risk of myocardial infarction and cardiac death based on clinical risk, coronary calcium, and epicardial adipose tissue: a prospective study. *Cardiovascular Research*, December 2019. doi:10.1093/cvr/cvz321. Https://academic.oup.com/cardiovascres/advance-article/doi/10.1093/cvr/cvz321/5680420.

[151] LeCun Y, Bengio Y, and Hinton G. Deep learning. *Nature*, 521(7553):436–444, May 2015. ISSN 1476-4687.

[152] Fu Y, Jung AW, Torne RV, Gonzalez S, Vöhringer H, Jimenez-Linan M, Moore L, and Gerstung M. Pan-cancer computational histopathology reveals mutations, tumor composition and prognosis. *bioRxiv*, page 813543, October 2019. doi:10.1101/813543. Publisher: Cold Spring Harbor Laboratory Section: New Results.

[153] Kather JN, Heij LR, Grabsch HI, Kooreman LFS, Loeffler C, Echle A, Krause J, Muti HS, Niehues JM, Sommer KAJ, Bankhead P, Schulte JJ, Cipriani NA, Ortiz-Brüchle N, Patnaik A, Srisuwananukorn A, Brenner H, Hoffmeister M, Brandt PAvd, Jäger D, Trautwein C, Pearson AT, and Luedde T. Pan-cancer image-based detection of clinically actionable genetic alterations. *bioRxiv*, page 833756, November 2019. doi:

10.1101/833756. Publisher: Cold Spring Harbor Laboratory Section: New Results.