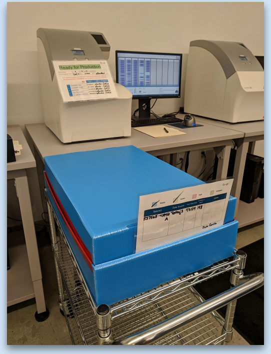
Digitizing Pathology Slides for ML Applications: Opportunities and Lessons Learned

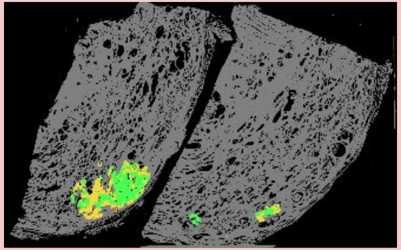
Craig Mermel, MD/PhD
Staff Research Scientist, Google Health

Overview

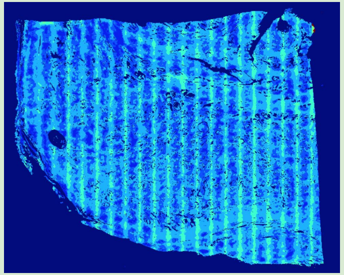
Digitization of Archival Slides



Machine Learning Applications in Pathology

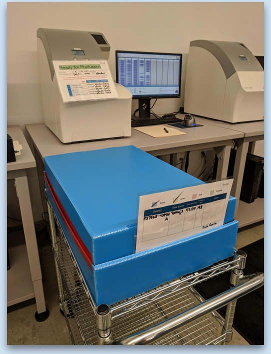


Impact of Image Quality on Machine Learning Applications



Overview

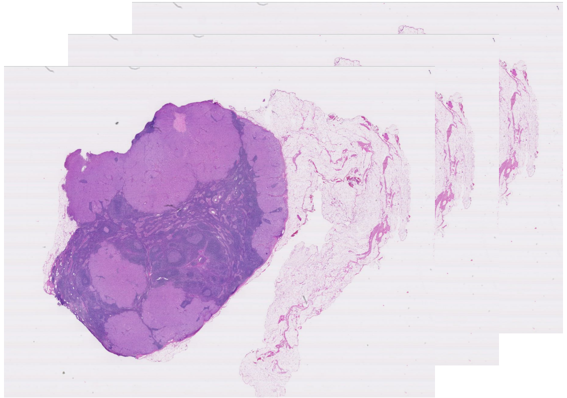
Digitization of Archival Slides



Machine Learning Applications in Pathology

Impact of Image Quality on Machine Learning Applications

Pathology Slide Digitization



~10 GB/slide
(40x magnification)

~150 years of

Pathology Knowledge

~10 MB
(raw text)

Aim is to **accurately digitize** the **vast amount** of **visual information** on glass slides at **high resolution**, for both **clinical** and **non-clinical** applications

Clinical vs Research Digitization

While the process of digitization is similar between clinical and research applications, the challenges are quite different

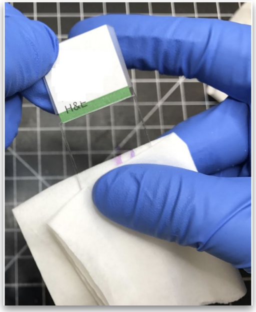
	Clinical Use	Non-Clinical/Research Use
Scale	Hundreds to thousands of slides (per day)	Hundreds of thousands to millions of slides
Latency	Low Latency Required	Medium to High Latency Often Acceptable
Input Material	Fresh cut & stained slides	Archival slides (years to decades old)
Tolerance to artifacts	Moderate (so long as doesn't interfere with diagnosis), can revert to glass	Low (small artifacts can hinder ability to train or validate models)
Slide Label	Digitize w/ PHI	Need to de-identify but preserve essential meta-data
Linkage to Clinical Data	Essential for proper diagnosis	Not essential but enabling for research

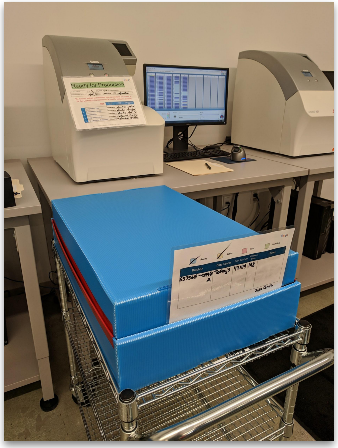
We primarily digitize archival (>10 yr old) slides

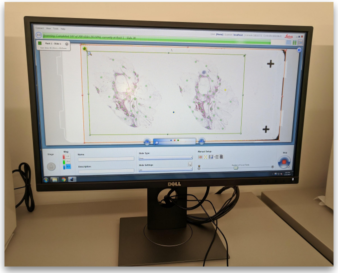
Slides are...

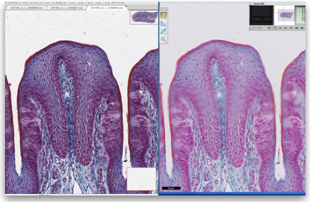
Images are...

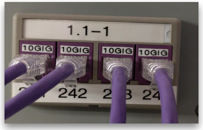
RECEIVED → RECORDED → RESTORED → SNAPSHOTTED → SETUP → SCANNED → QC'D & UPLOADED







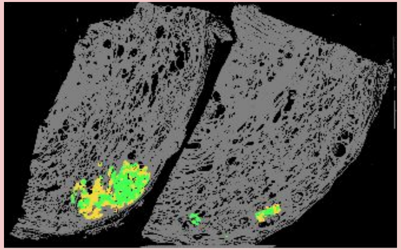




Overview

Digitization of Archival Slides

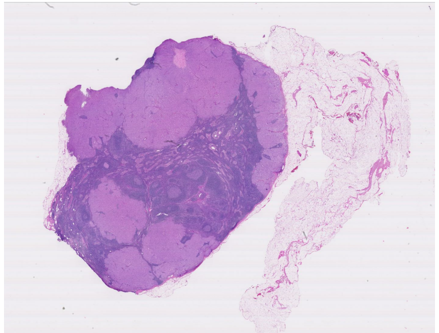
Machine Learning Applications in Pathology



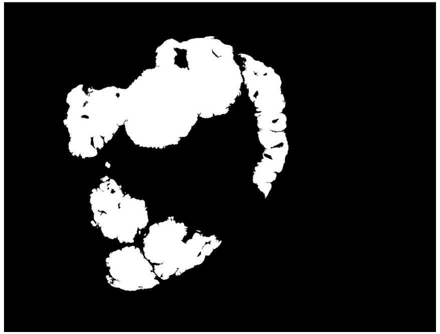
Impact of Image Quality on Machine Learning Applications

Strongly supervised learning vs Weakly supervised learning

Strongly supervised learning

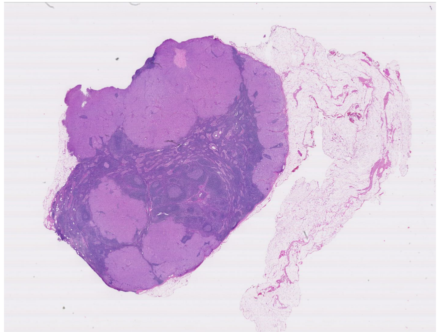


+



- WSI + segmentation mask
- annotations costly
- each WSI is ~100k of samples
- 1000s WSIs usually sufficient*

Weakly supervised learning



+

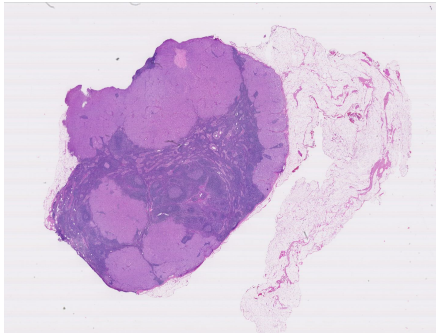
"has tumor"
"grade III breast cancer"
"5y+ survival"

- WSI + slide/case label
- annotations cheap
- each WSI is one sample
- 1000s WSIs usually not sufficient*
- can detect features unknown to pathologist

* including generalization to other slide sources, scanners, demographics, etc

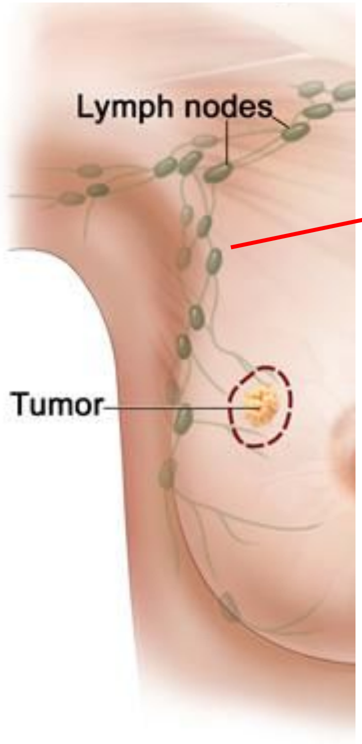
Unsupervised learning

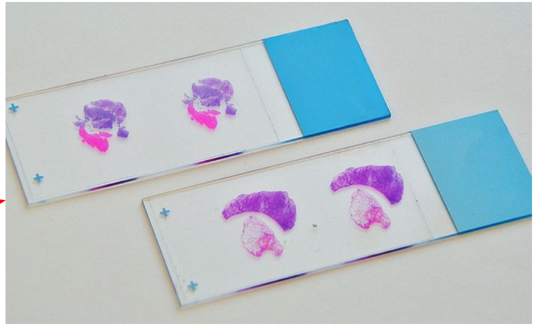
Unsupervised learning

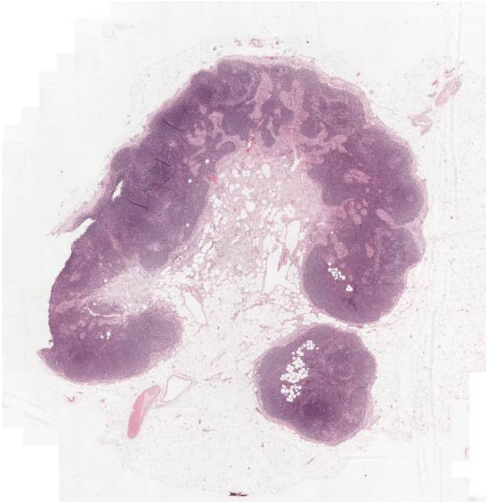
 +

- Learns from unlabelled images
- Find patterns/structure in data “unbiased” by prior pathology knowledge
- Often need labelled data to understand the results

Detecting metastases in lymph nodes is important for tumor staging



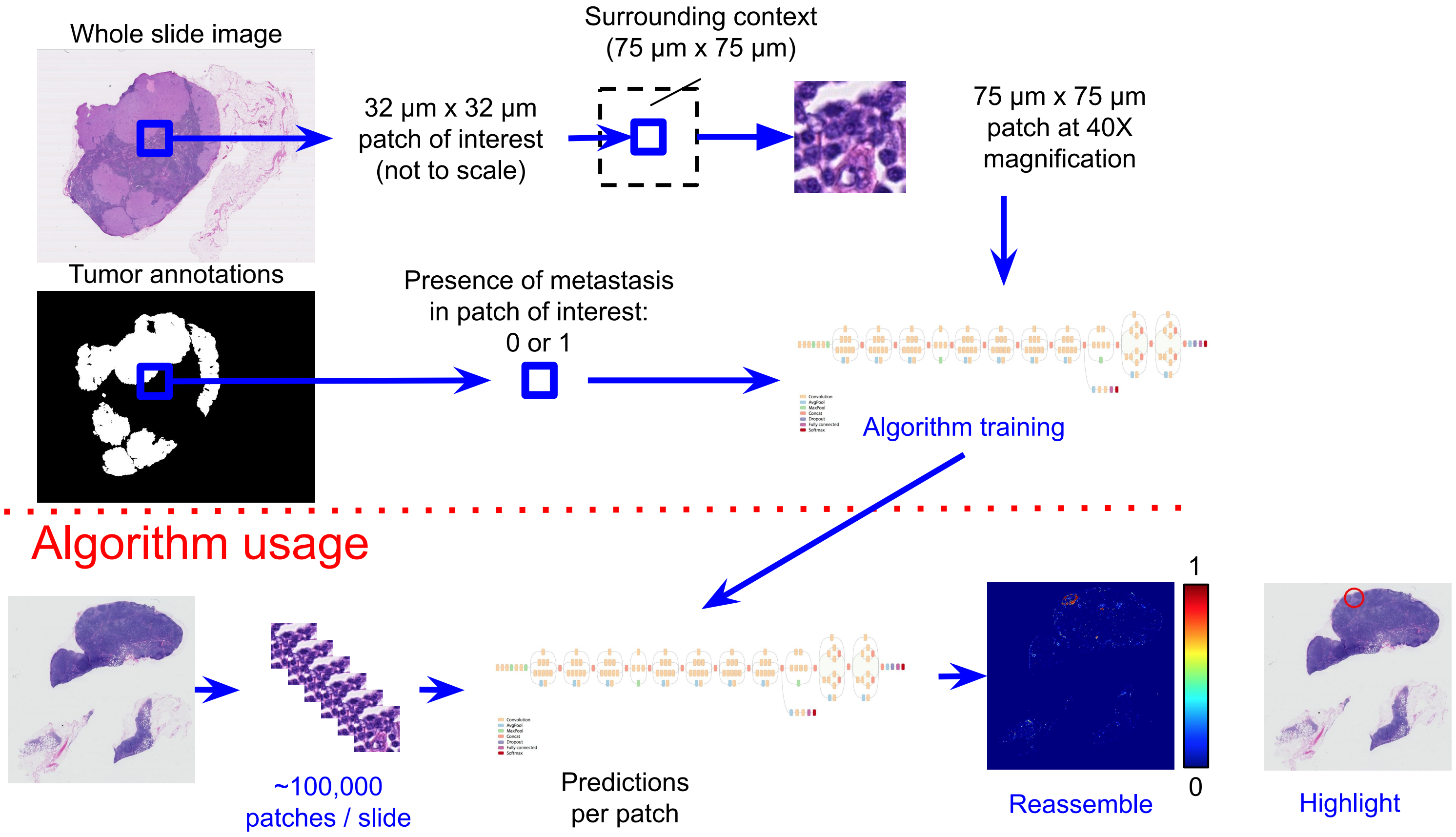




Sentinel lymph node status in breast cancer:
- Informs prognosis and therapy decisions

image from www.medicinebtg.com

Training an AI algorithm to detect metastases in lymph nodes



Performance in tumor localization - Camelyon16 challenge data set

Tumor localization score (FROC):

- Single pathologist: 0.73*
- Camelyon16 winner: 0.81
- Google AI algorithm: 0.91

The algorithm also generalizes to data from other clinics and scanners

** unlimited time (30h), but 0 false positives*

Artificial Intelligence–Based Breast Cancer Nodal Metastasis Detection

Insights Into the Black Box for Pathologists

Yun Liu, PhD; Timo Kohlberger, PhD; Mohammad Norouzi, PhD; George E. Dahl, PhD; Jenny L. Smith, MD; Arash Mohtashamian, MD; Niels Olson, MD; Lily H. Peng, MD, PhD; Jason D. Hipp, MD, PhD; Martin C. Stumpe, PhD

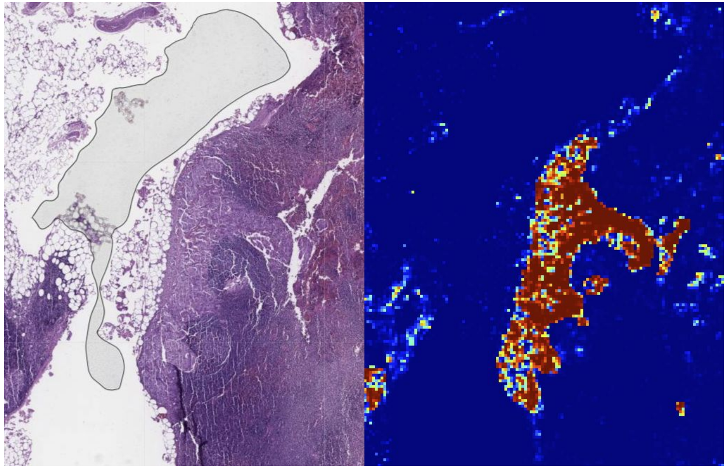
Slide level AUC:

- **Single pathologist: 96.6%***
- **Google AI Algorithm: 99.3%**

Liu et al, Archives of Pathology, 2018

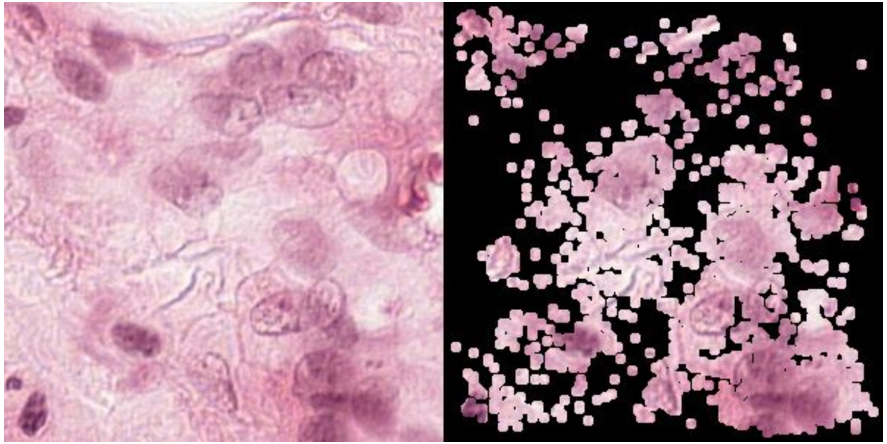
An independent clinical data set scanned with different scanners showed similar FROC results

True positive



Accurate despite: Air bubble, cutting artifacts, hemorrhagic, and necrotic and poorly processed tissue

False positive

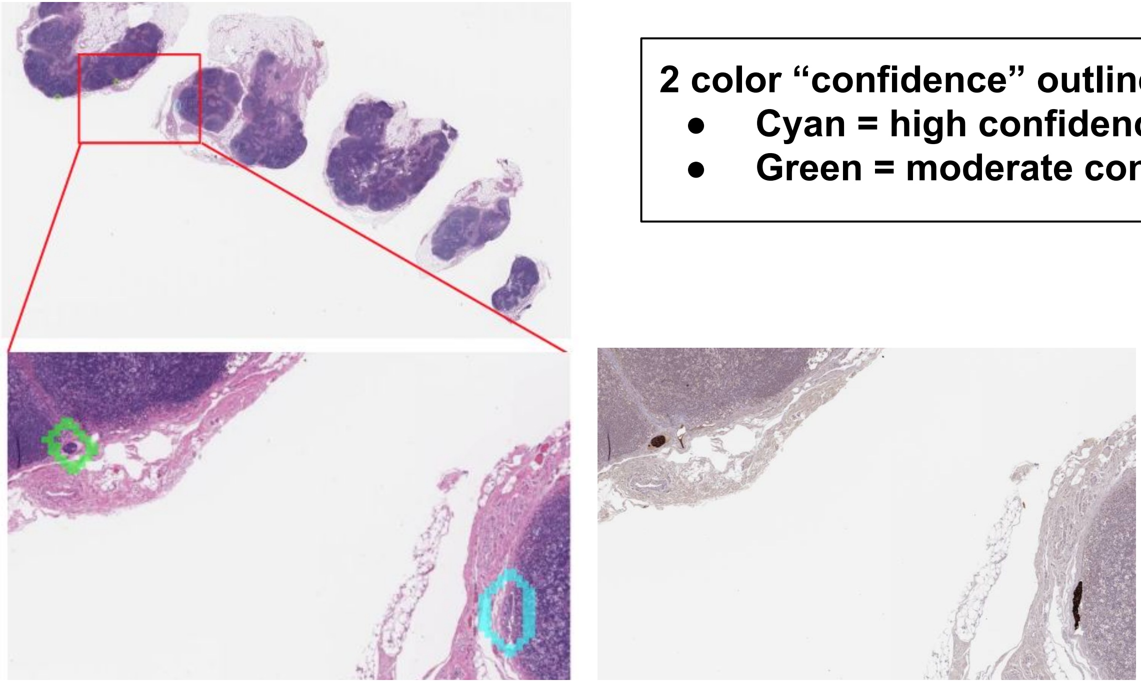


Large out of focus, overlapping histiocytes

Liu et al, Archives of Pathology, 2018

Evaluating the utility of the AI algorithm to pathologists

Hypothesis: The lymph node AI algorithm can improve the efficiency of pathologists



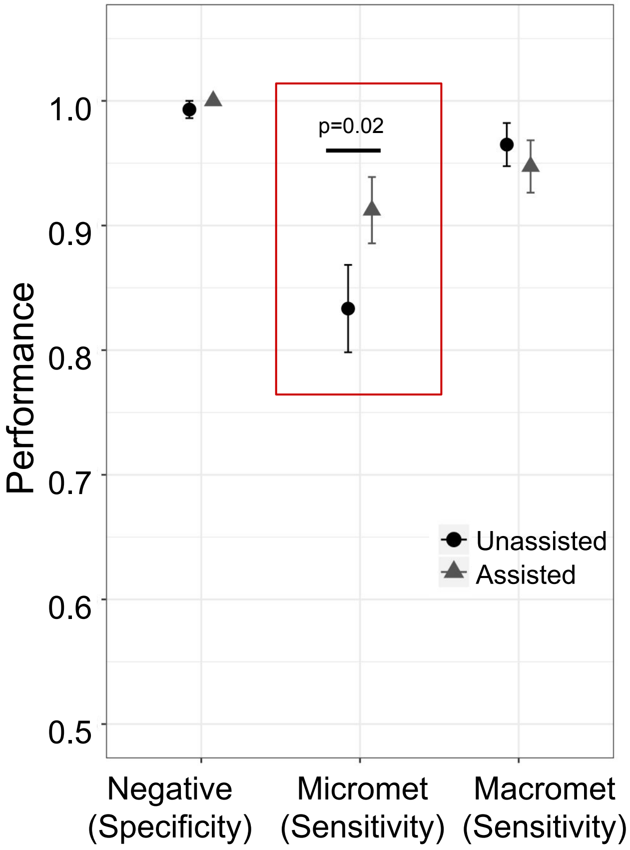
2 color "confidence" outlines:

- Cyan = high confidence (high specificity)
- Green = moderate confidence (high sensitivity)

IHC (cytokeratin stain)

Steiner et al, AJSP, 2018

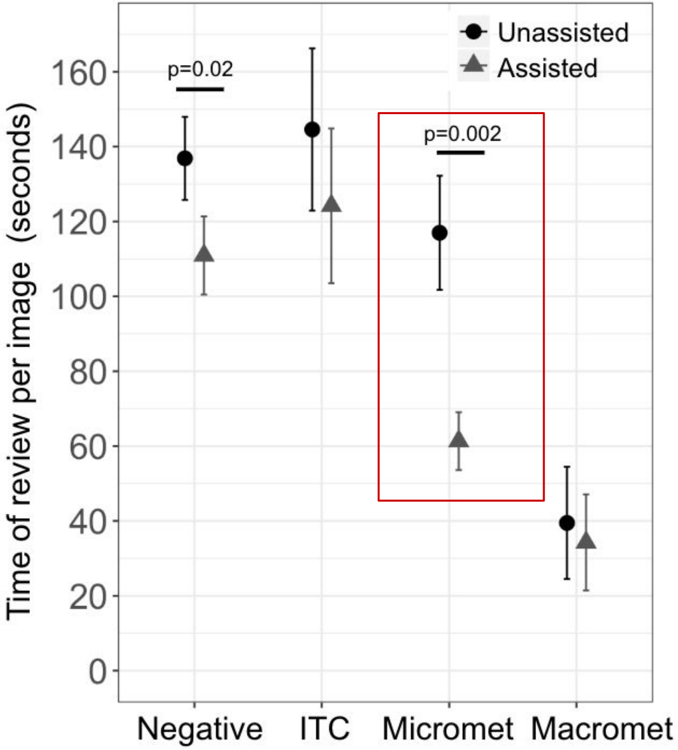
The AI algorithm improved accuracy of tumor detection



Accuracy (micromets):
With assistance: 91%
Without assistance: 83%
→ error reduced by ~½

Steiner et al, AJSP, 2018

The AI algorithm improved pathologist efficiency



Time benefit:

- Negative: 111 vs. 137 s (*P=0.018)*
- Micromets: 61 vs. 116 s *(P=0.002)*

Steiner et al, AJSP, 2018

Prostate cancer Gleason grading

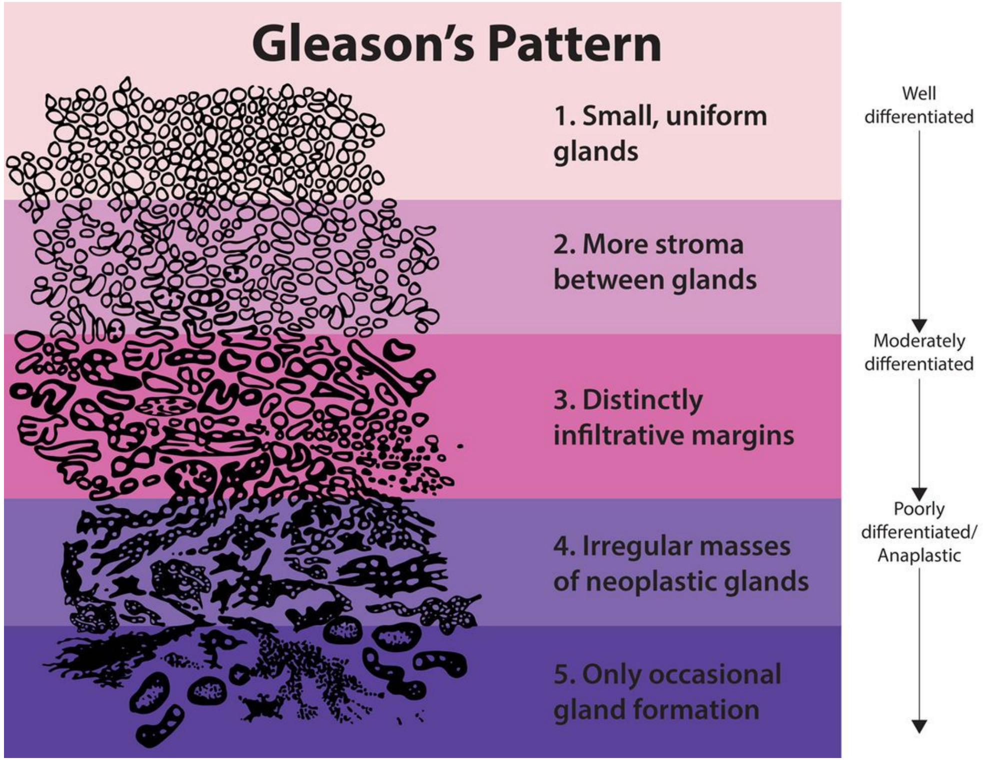


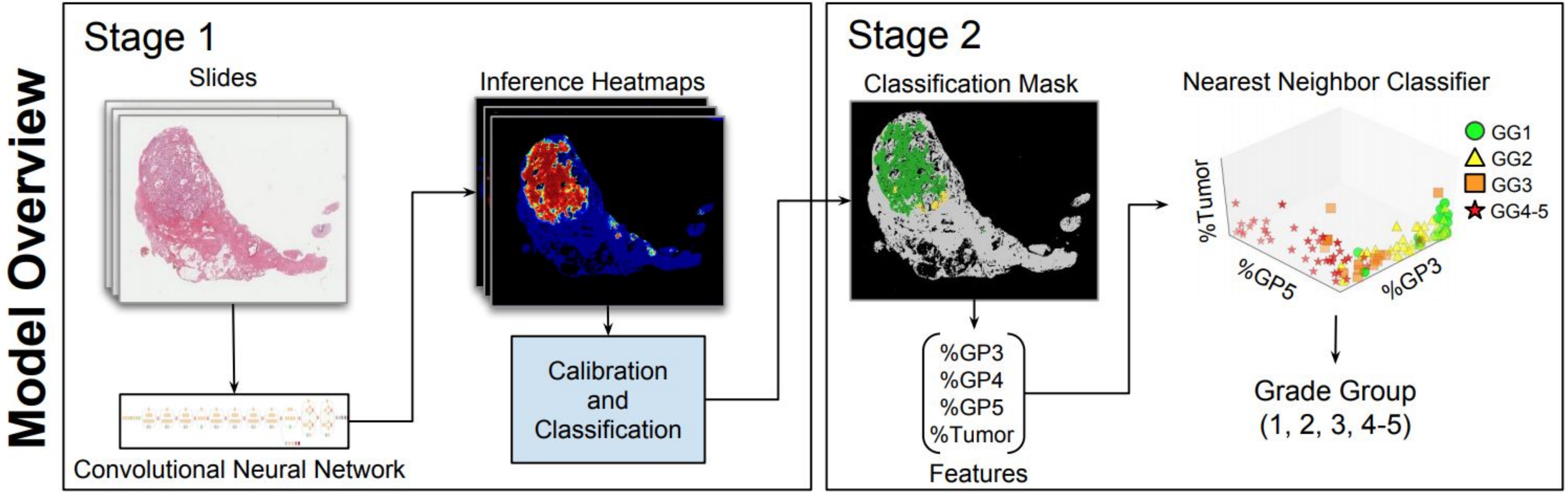
image credit: Wikipedia

- 2nd most common cancer in men (in North America)
- Gleason grade has direct impact on treatment decision
- highly subjective classification task, large intergrader variability

A Deep Learning System For Gleason grading

Two-stage model:

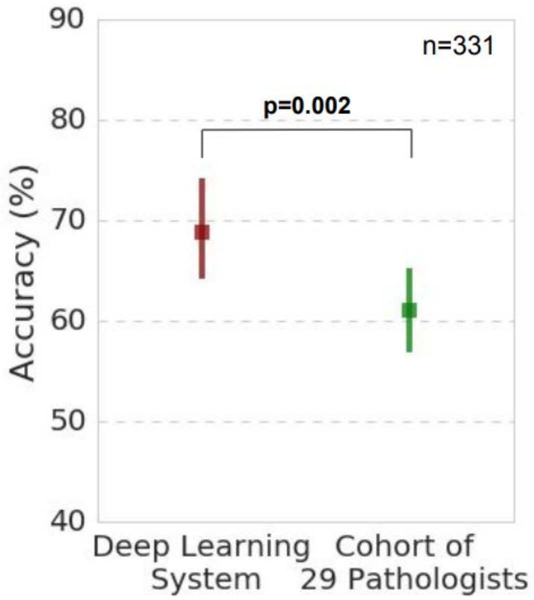
1. Local Gleason classification
2. Slide summarization



Nagpal et al, *npj Digital Medicine,* June 2019

Our Gleason grading model outperforms general pathologists in on radical prostatectomy specimens

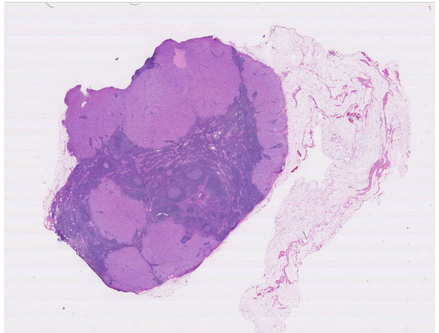
Radical Prostatectomies



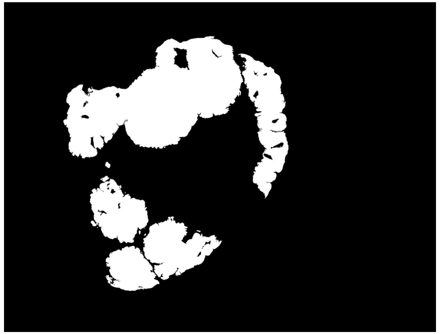
Nagpal et al, *npj Digital Medicine,* June 2019

Strongly supervised learning vs Weakly supervised learning

Strongly supervised learning

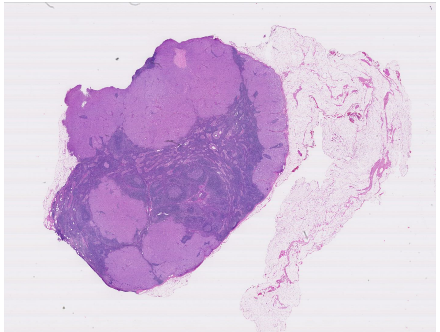


+



- WSI + segmentation mask
- annotations costly
- each WSI is ~100k of samples
- 1000s WSIs usually sufficient*

Weakly supervised learning

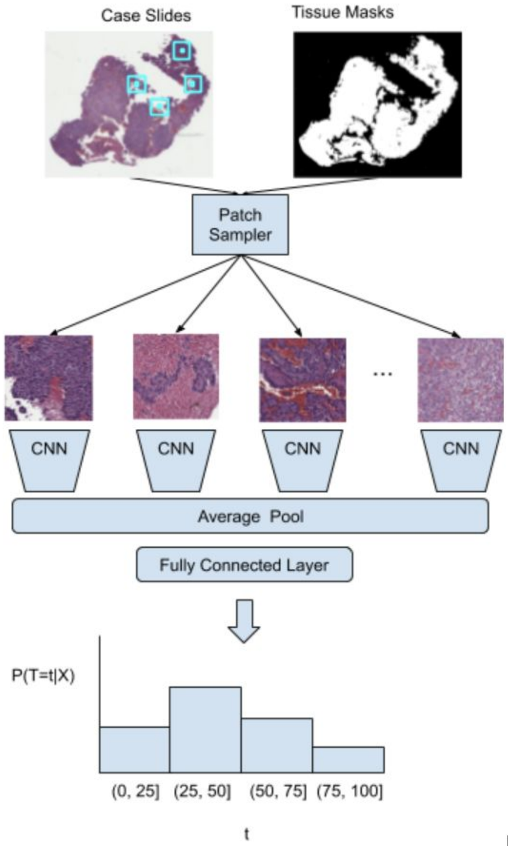


+ "has tumor"
"grade III breast cancer"
"5y+ survival"

- WSI + slide/case label
- annotations cheap
- each WSI is one sample
- 1000s WSIs usually not sufficient*
- can detect features unknown to pathologist

* including generalization to other slide sources, scanners, demographics, etc

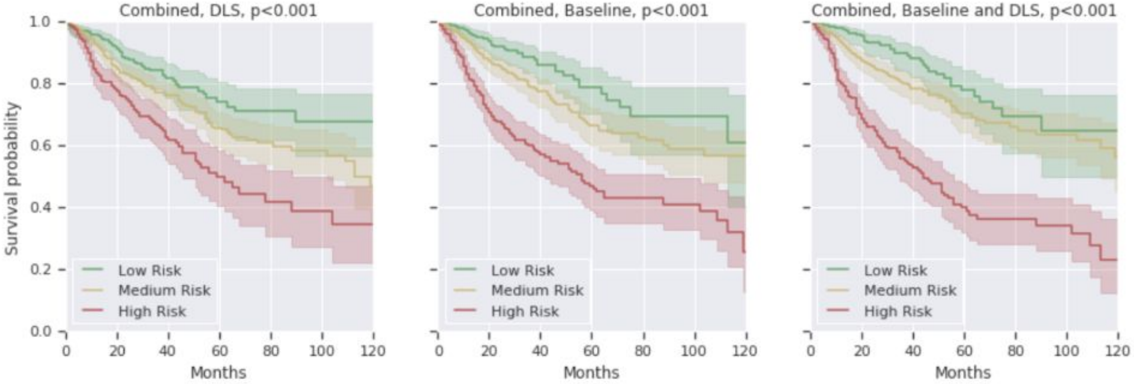
Weakly supervised learning: Direct survival prediction



DLS only

Age + Gender+ Stage

DLS + Age + Gender + Stage

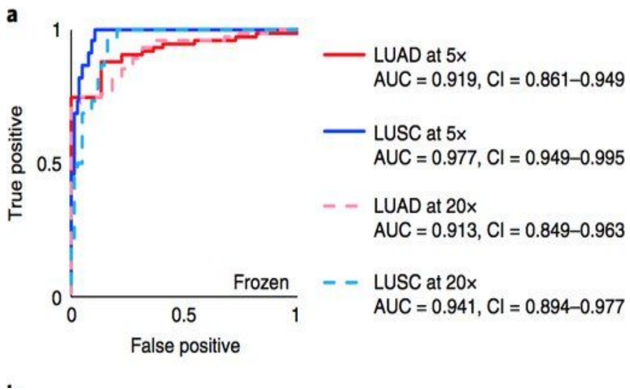


Pan cancer analysis (4,880 across 10 TCGA cancer types)

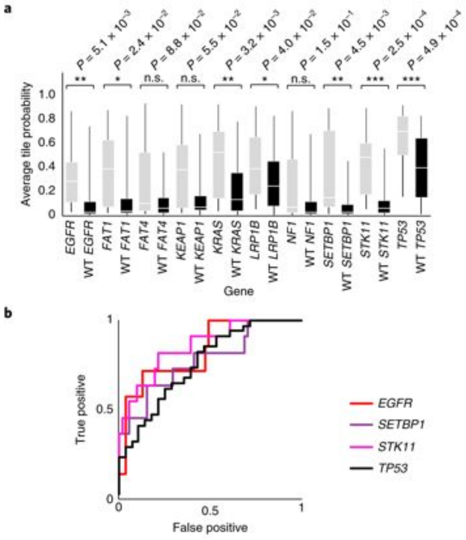
Wulczyn et al, https://arxiv.org/abs/1912.07354

Other Weak Label Examples (non-Google)

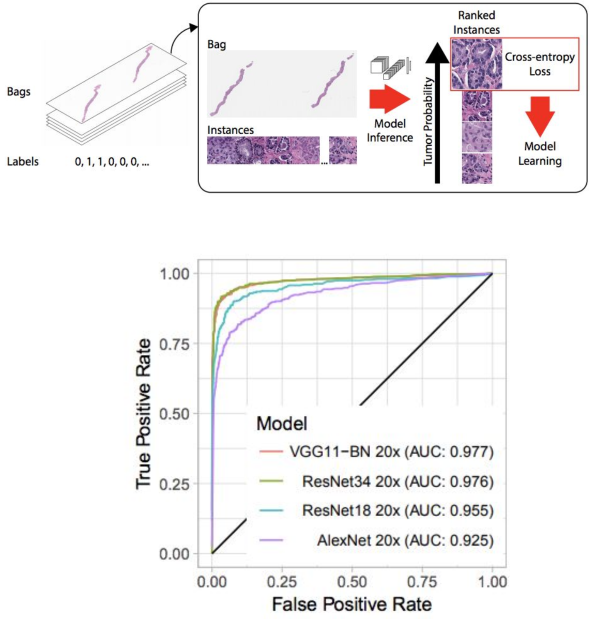
Lung Adeno vs. Squamous Weak Label Prediction



Direct Prediction of Mutations from H&E Images



Slide Level Prostate Cancer Detection in Needle Core Biopsies

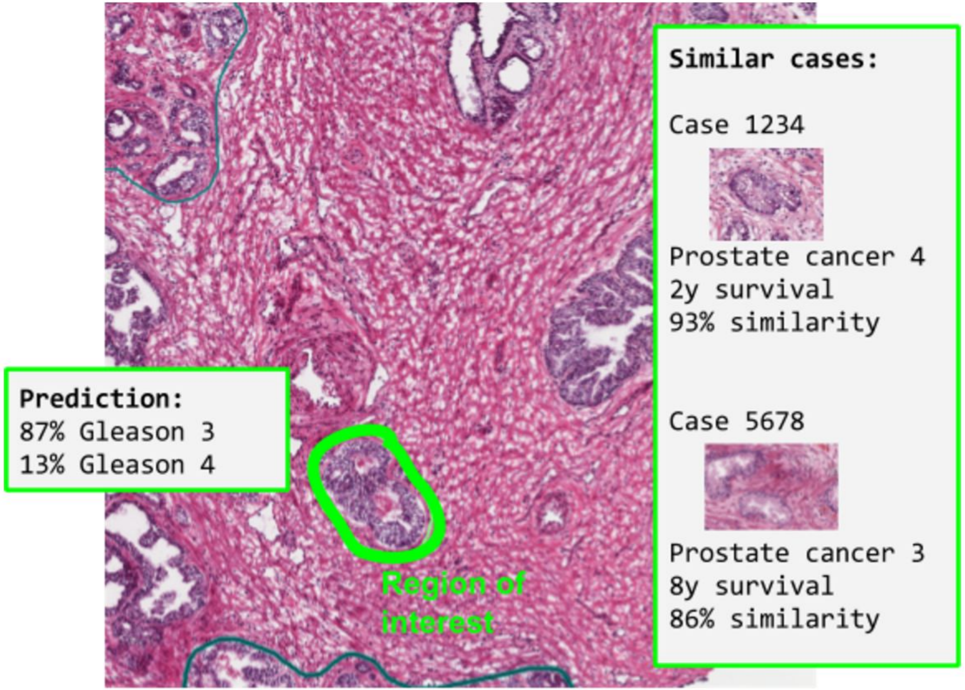


Coudray et al, Nature Medicine (Oct 2018)

Campanella et al, *Nature Medicine*, Aug 2019

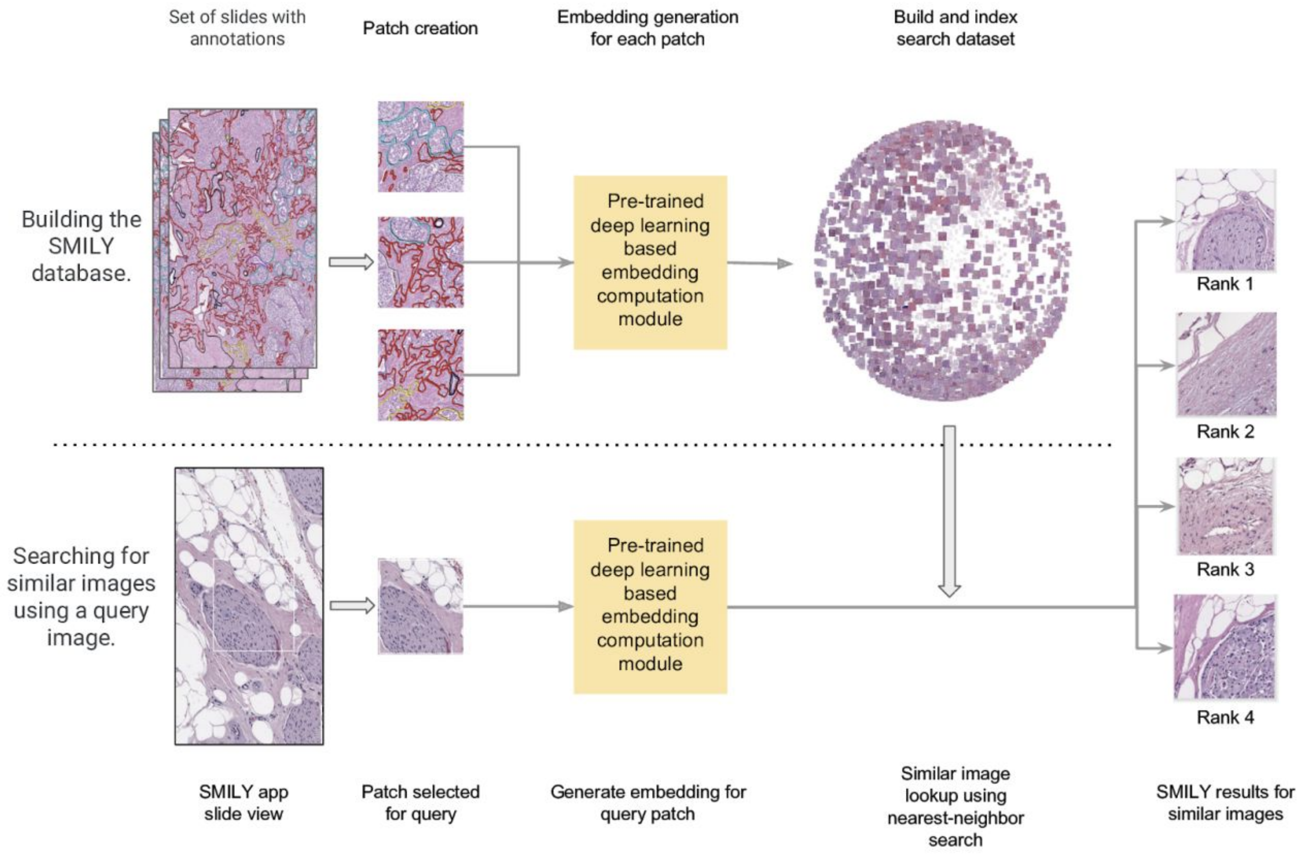
Unsupervised Similar Image Search for Pathology (SMILY)

Smart archival lookup tool for pathologists to find cases that are visually similar



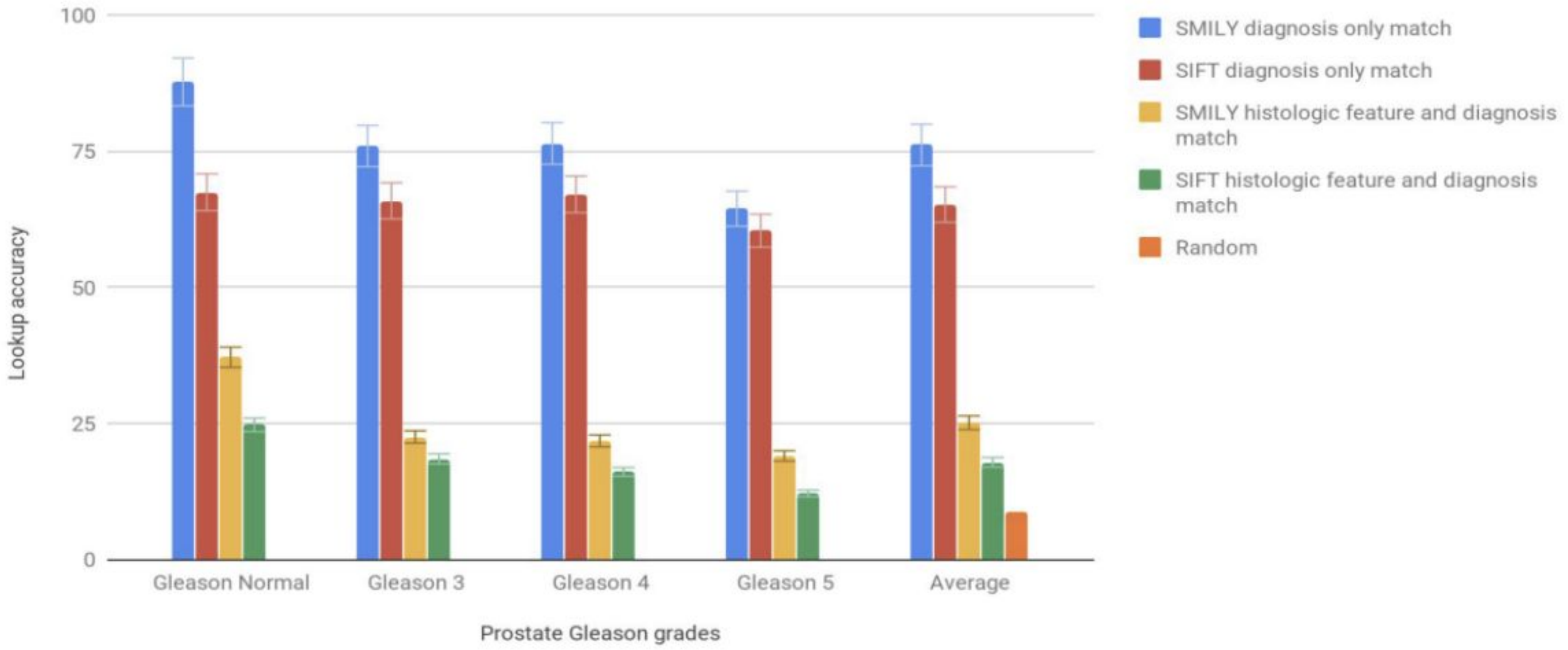
Mock of the SMILY panel (right) in the pathology frontend along with the classification predictions (left).

Similar Image Search for Pathology (SMILY)



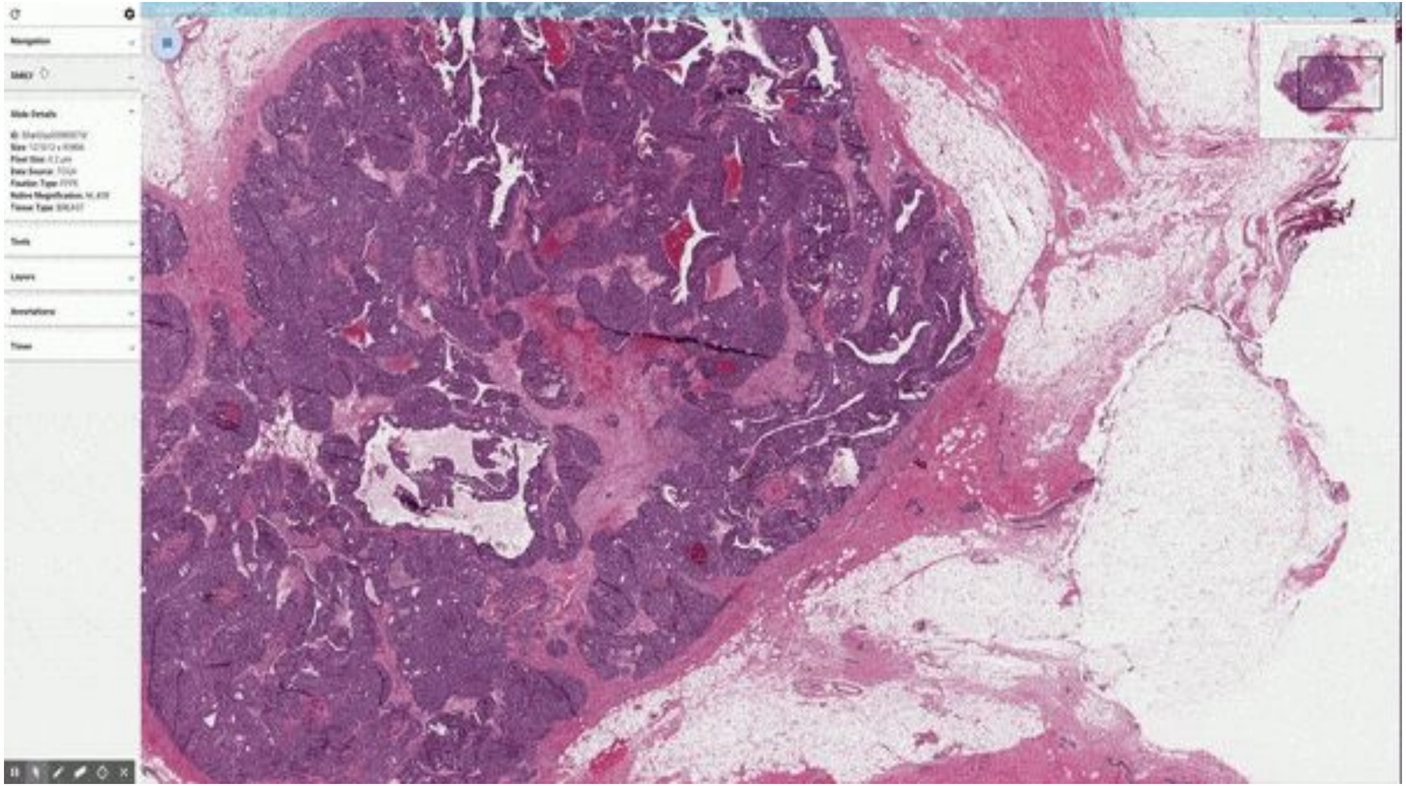
Hegde et al, npj Digital Medicine, June 2019

Similar Image Search for Pathology (SMILY)



Hegde et al, npj Digital Medicine, June 2019

Similar Image Search for Pathology (SMILY)



Hegde et al, npj Digital Medicine, June 2019

Overview

Digitization of Archival Slides

Machine Learning Applications in Pathology

Impact of Image Quality on Machine Learning Applications

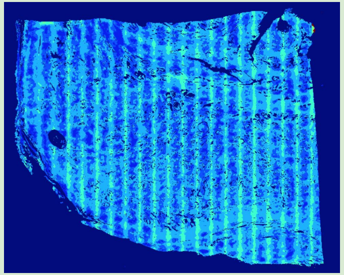
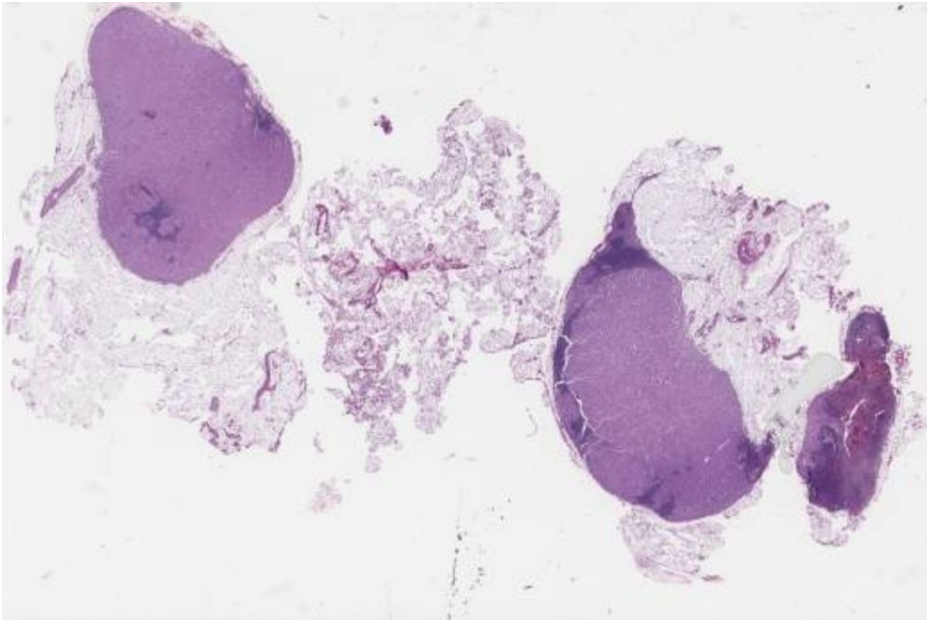
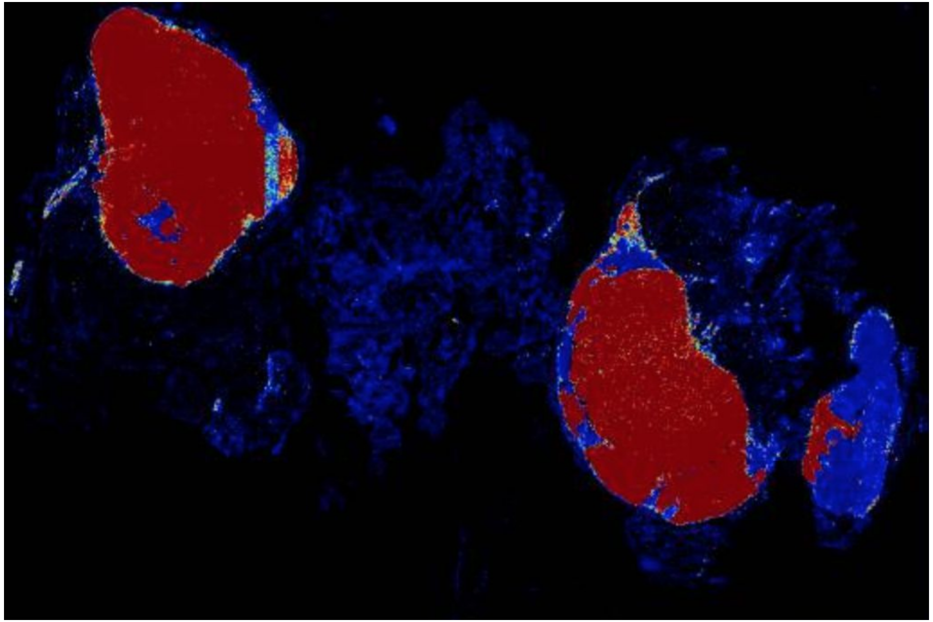


Image quality matters

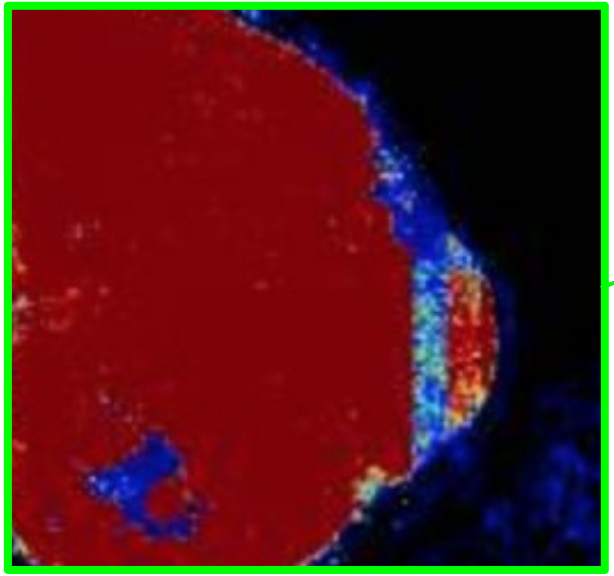


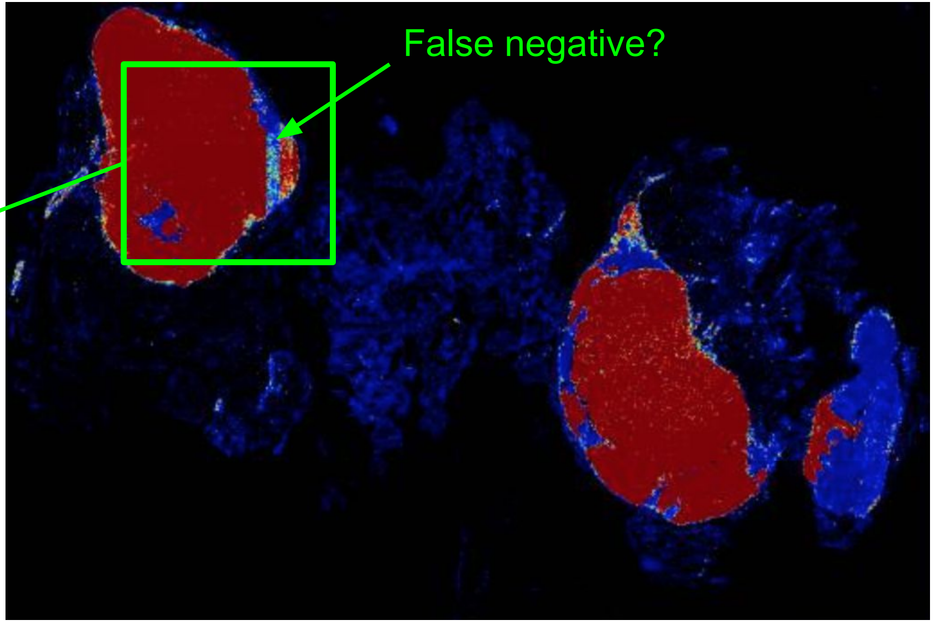
Lymph node biopsy



Tumor prediction

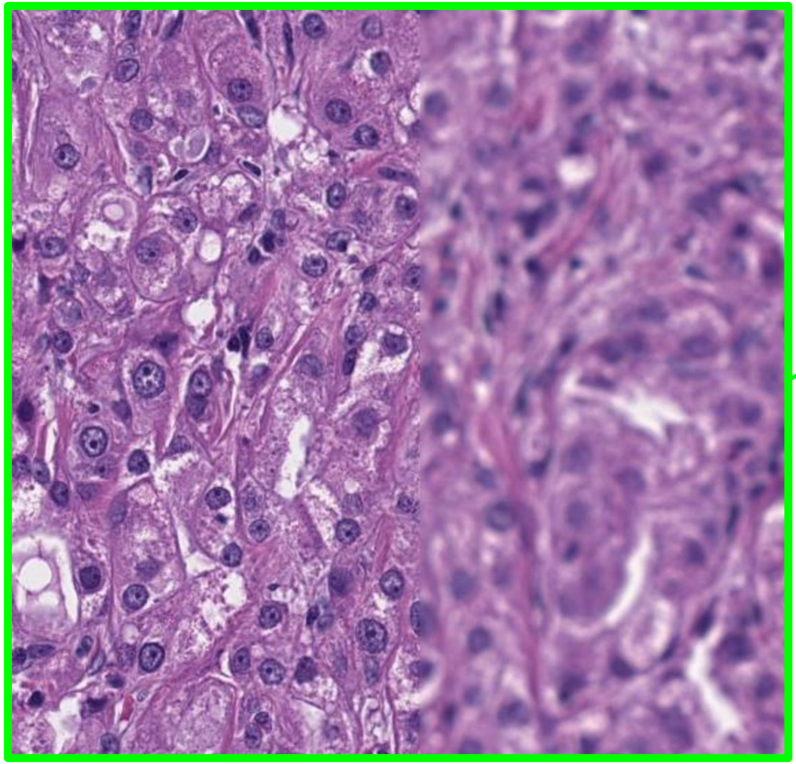
Image quality matters

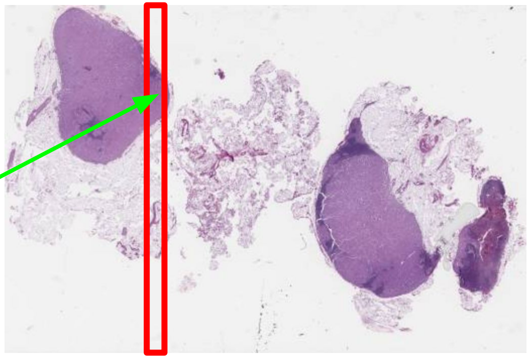




Tumor prediction

Image quality matters





- Entire scan column out of focus
- Confuses model (and potentially pathologist)
- Mitigation: detect and flag ungradable areas

Training a Classifier to Detect OOF Patches

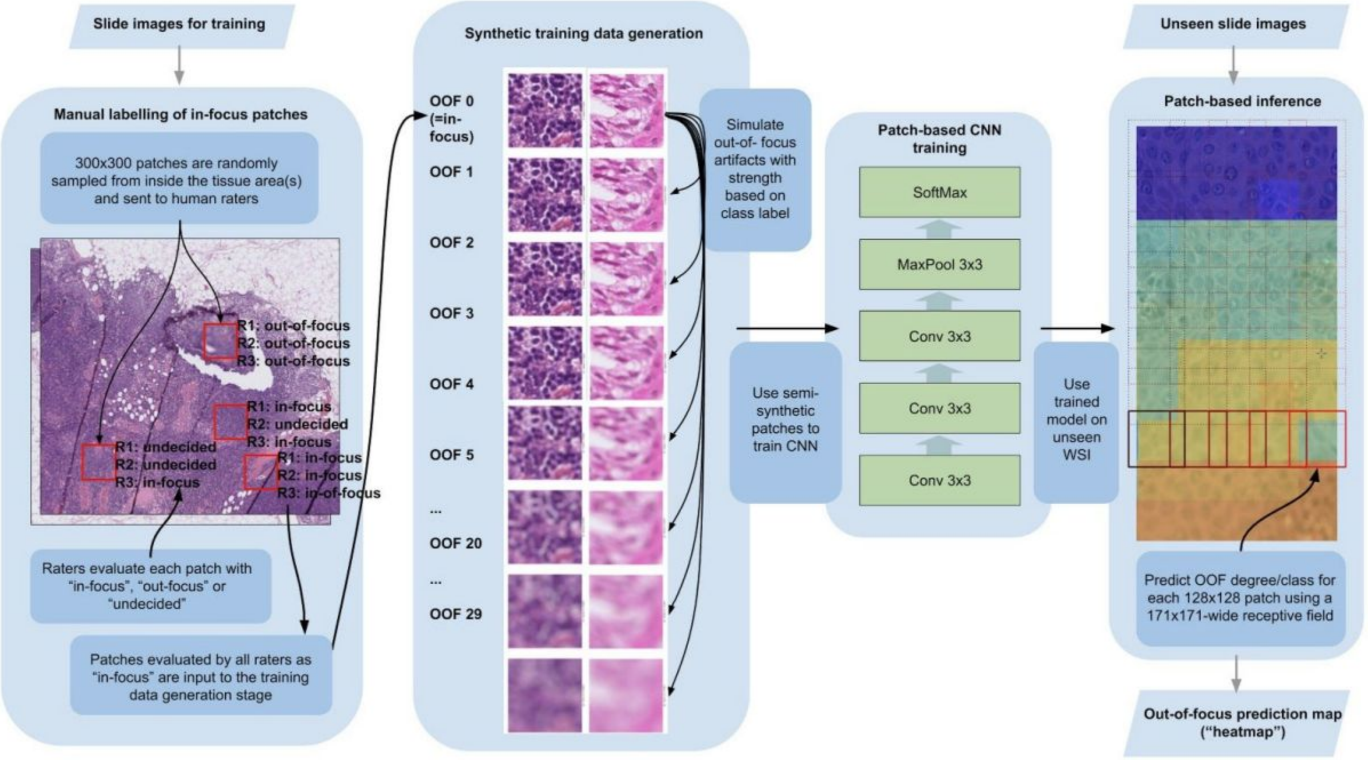
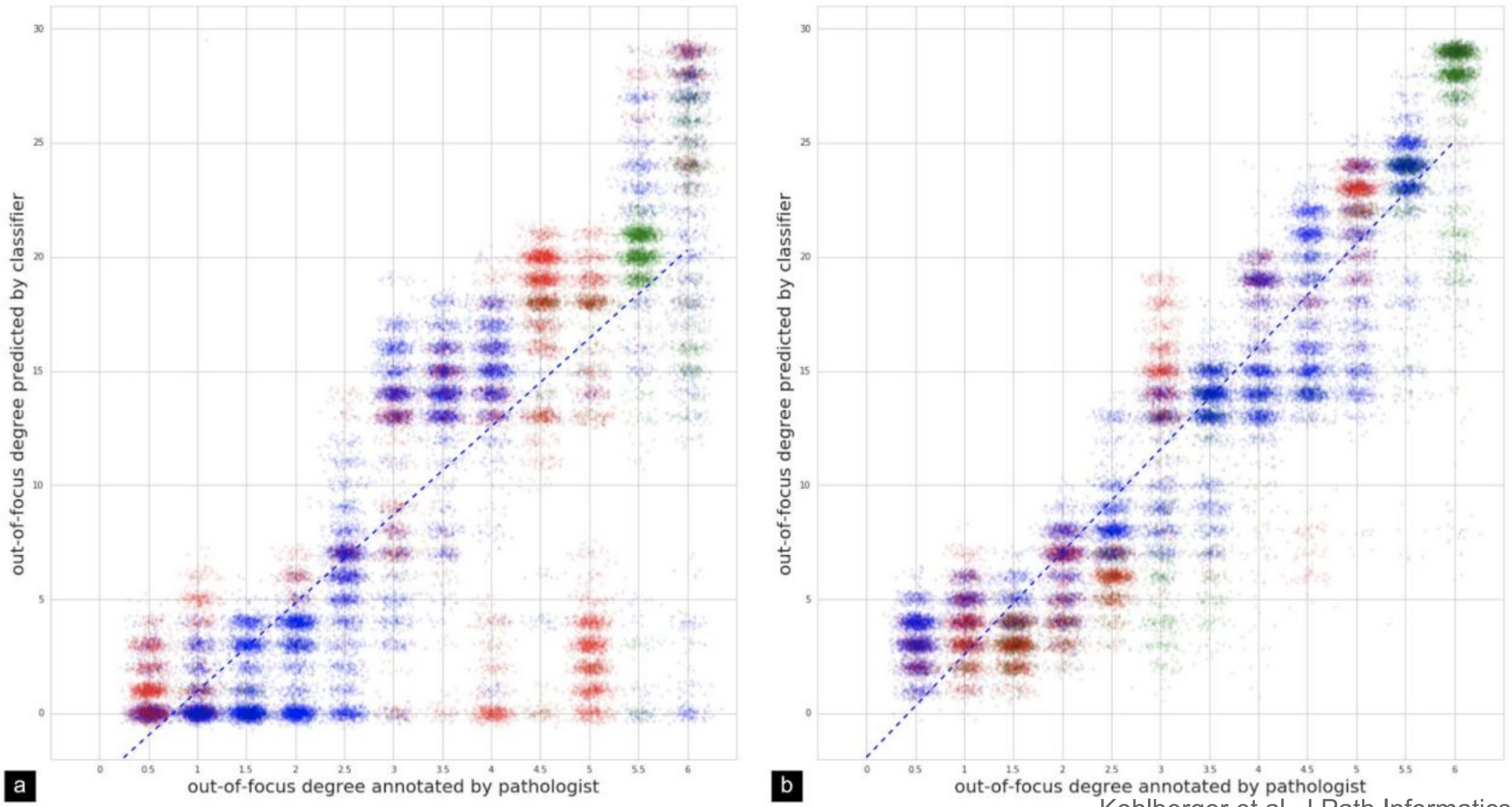


Figure 2: Overview of our convolutional neural network (CNN) approach to automated out-of-focus (OOF) grading: ConvFocus.

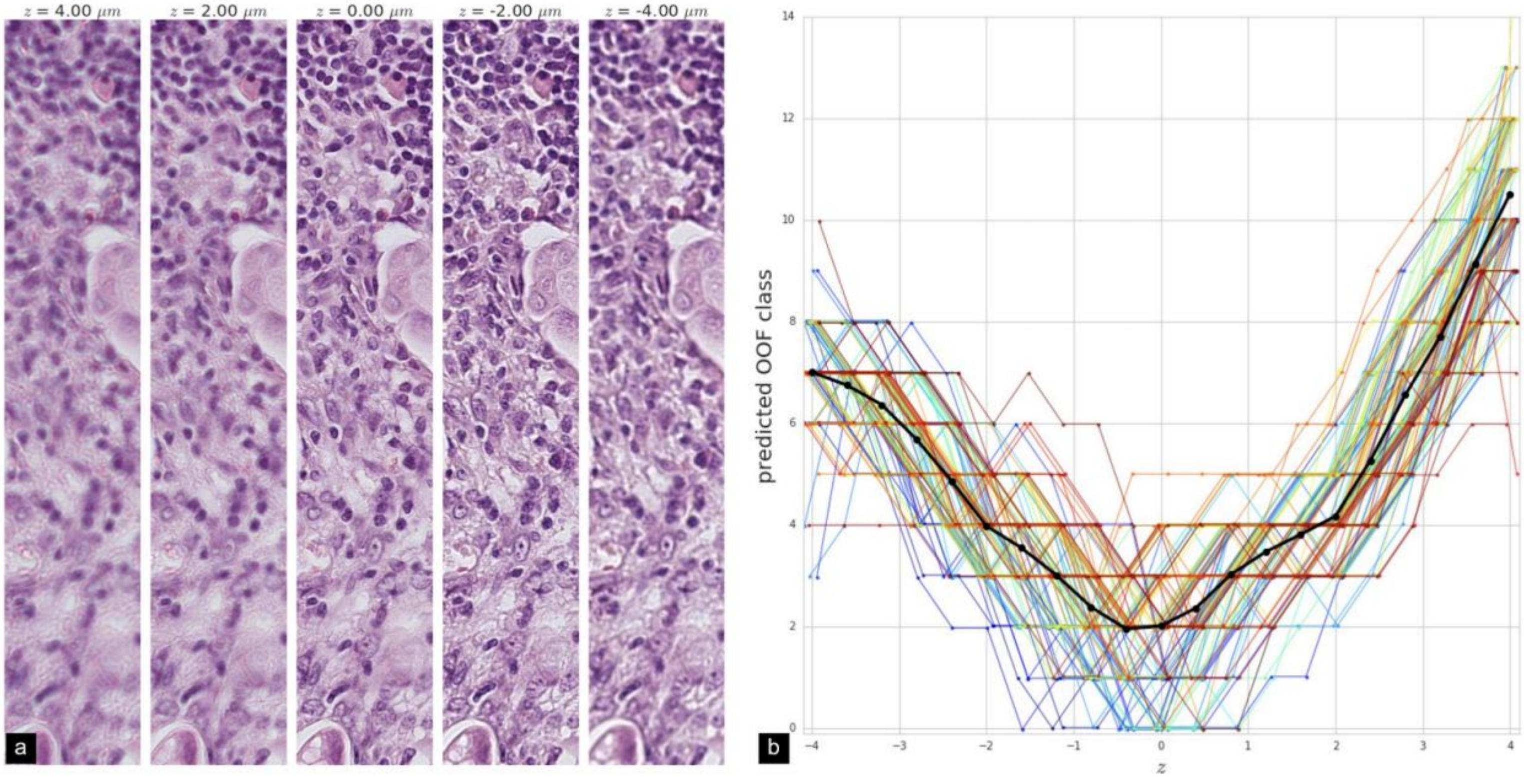
Kohlberger et al, J Path Informatics, Dec 2019.

Predicted Focus Quality vs. Pathologist Annotation for 2 Different Scanner Types



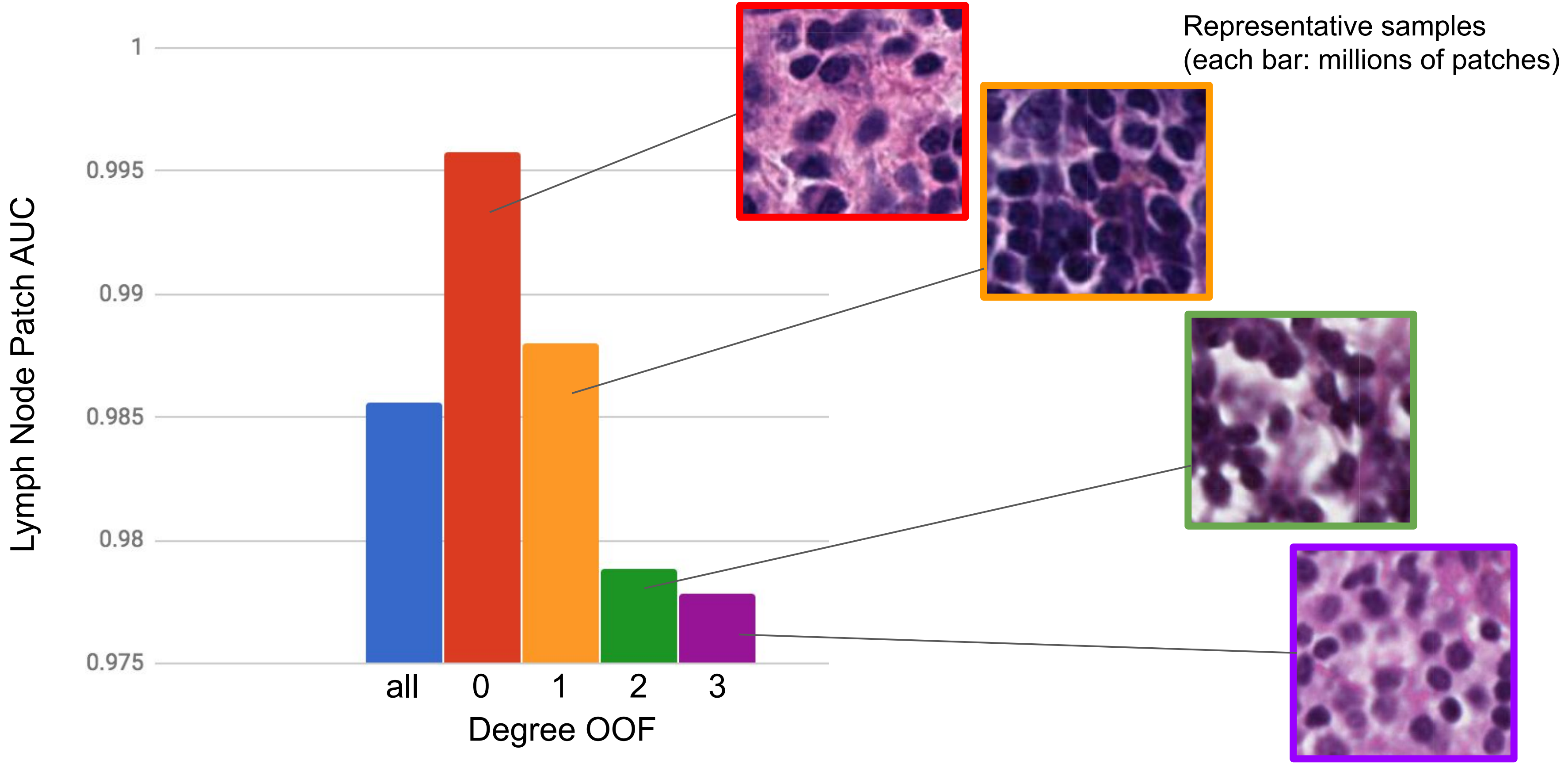
Kohlberger et al, J Path Informatics, Dec 2019.

OOF class vs. z-stack depth



Kohlberger et al, J Path Informatics, Dec 2019.

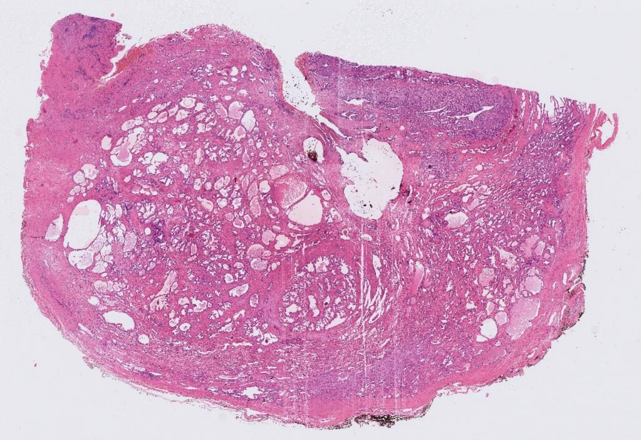
How image quality impacts model performance



Kohlberger et al, J Path Informatics, Dec 2019.

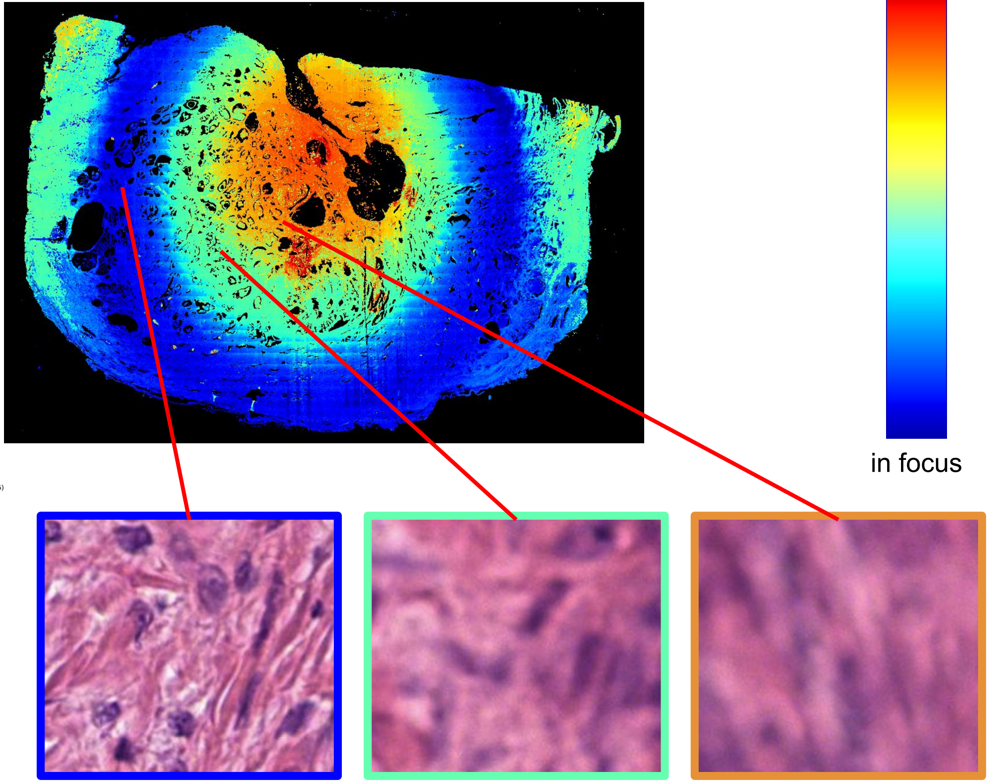
Automatic quality control for all images

digitized slide



focus
classifier

focus quality map



out of focus

in focus