Bias & Batch Effects in Medical Imaging

Towards fair and robust solutions

Heather Couture

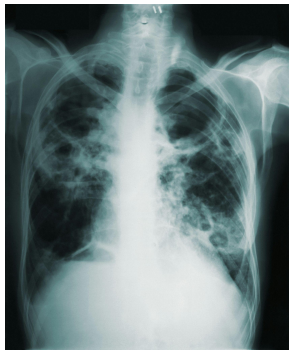
March 15, 2025
11 am EDT

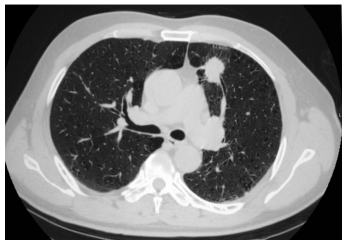
30 minutes + Q&A

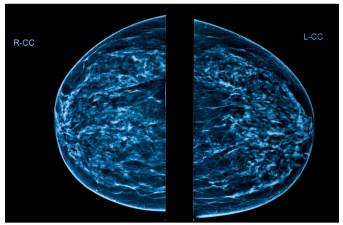
Pixel Scientia LABS

Radiology: Deep Learning Can Predict Race

Generally not possible for human experts





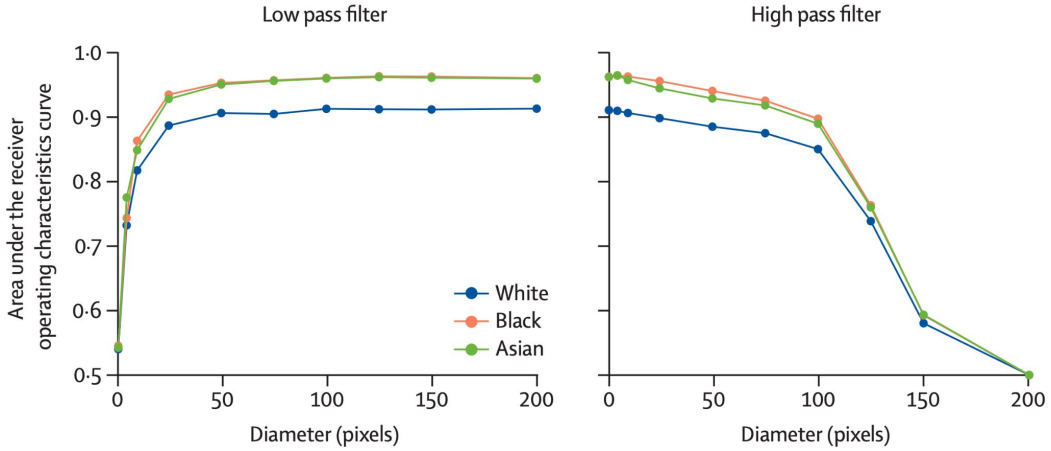


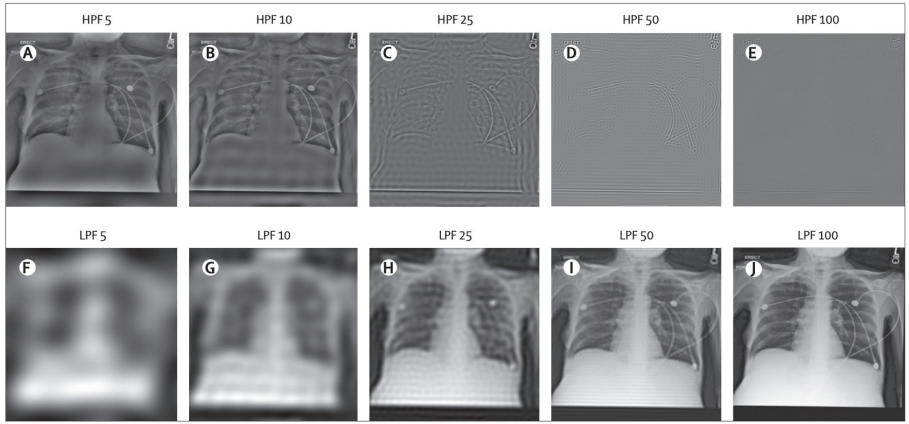
Source: Canva

	Area under the receiver operating characteristics curve
Race detection in radiology imaging	
Chest x-ray (internal validation)*	
MXR (Resnet34, Densenet121)	0·97, 0·94
CXP (Resnet 34)	0·98
EMX (Resnet34, Densenet121, EfficientNet-B0)	0·98, 0·97, 0·99
Chest x-ray (external validation)*	
MXR to CXP, MXR to EMX	0·97, 0·97
CXP to EMX, CXP to MXR	0·97, 0·96
EMX to MXR, EMX to CXP	0·98, 0·98
Chest x-ray (comparison of models)†	
MXR, CXP, EMX	Multiple results (appendix p 26)
CT chest (internal validation)*	
NLST (slice, study)	0·92, 0·96
CT chest (external validation)*	
NLST to EM-CT (slice, study)	0·80, 0·87
NLST to RSPECT (slice, study)	0·83, 0·90
Limb x-ray (internal validation)*	
DHA	0·91
Mammography*	
EM-Mammo (image, study)	0·78, 0·81
Cervical spine x-ray*	
EM-CS	0·92

Source: Gichoya, AI recognition of patient race in medical imaging: a modelling study, 2022

Even When the Images are Severely Corrupted

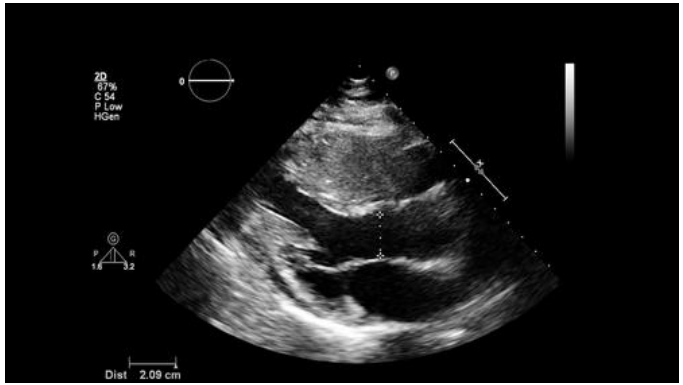




"[W]e emphasise that the ability of AI to predict racial identity is itself not the issue of importance, but rather that this capability is readily learned and therefore is likely to be present in many medical image analysis models, providing a direct vector for the reproduction or exacerbation of the racial disparities that already exist in medical practice."

Source: Gichoya, AI recognition of patient race in medical imaging: a modelling study, 2022

Cardiac Ultrasound: Confounding Variables → Shortcut



Source: Shutterstock

AUC on engineered dataset with race confounded by age or sex

Bias	Race prediction biased by sex	Race prediction biased by age
0.5	0.57	0.53
0.6	0.67	0.59
0.7	0.73	0.66
0.8	0.79	0.73
0.9	0.82	0.79
1.0	0.84	0.85

Source: Duffy, Confounders mediate AI prediction of demographics in medical imaging, 2022

Goals of This Webinar

How bias manifests in medical images

Case study on histopathology

Detection

Mitigation

Who am I?

- Heather Couture
- Computer vision consultant

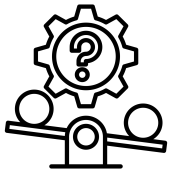
OWKIN DECIPHEX ULTIVUE AGENDIA Leica MICROSYSTEMS QRITIVE

- Keynote speaker at MICCAI workshop on computational pathology
- Contributor to Scientific American, The Pathologist, DPA Blog
- Newsletter and podcast

Computer Vision Insights by Pixel Scientia Labs

- PhD in Computer Science from University of North Carolina

Sources of Bias



Bias: systematic deviation from fairness or accuracy that leads to partial or prejudiced outcomes

Dataset bias: under-representation of demographic or geographic groups

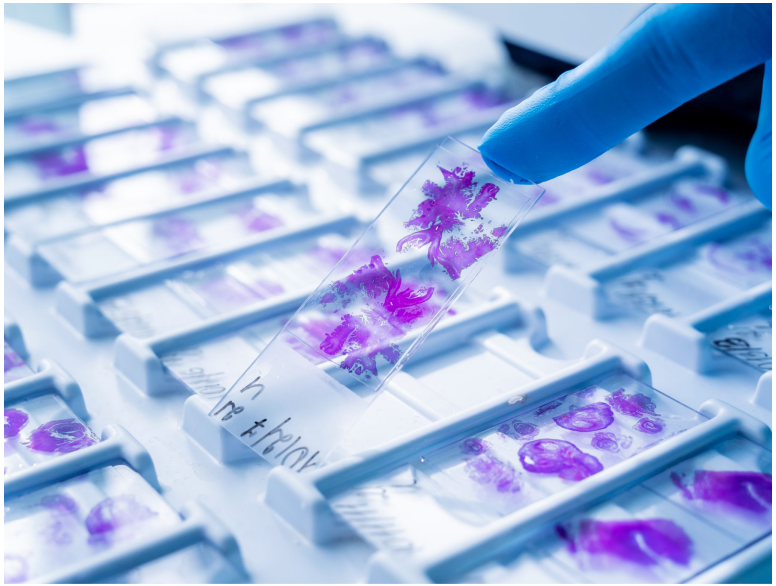
Sampling bias: class or medical center imbalance

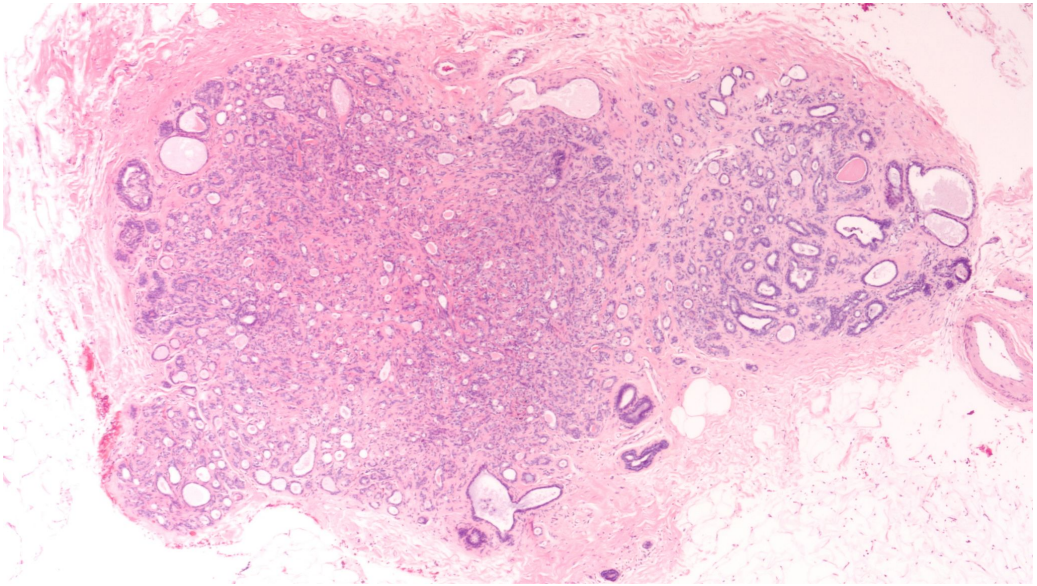
Technical bias: scanner and tissue preparation

Annotation bias: clinician subjectivity

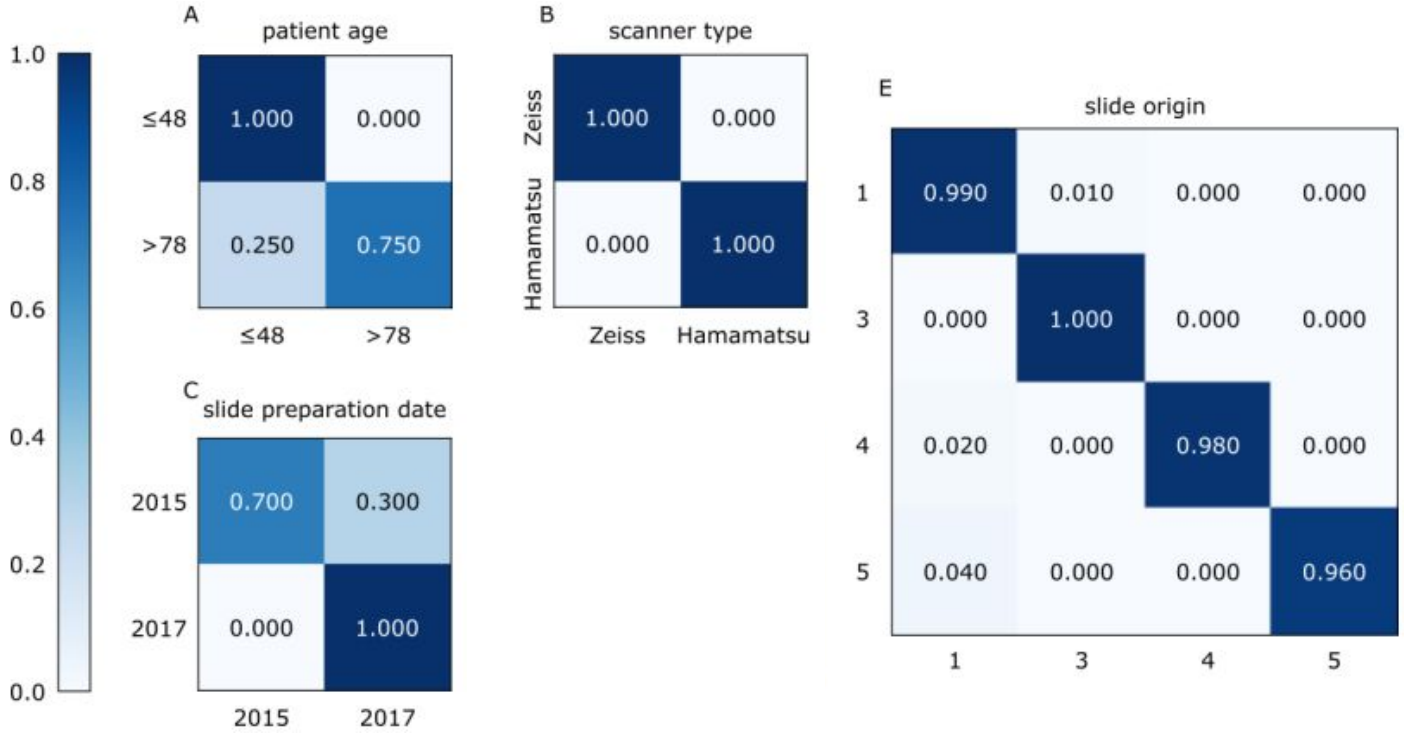
Modeling bias: feature extraction shortcuts, architecture decisions

Case Study: Sources of Bias in Histopathology





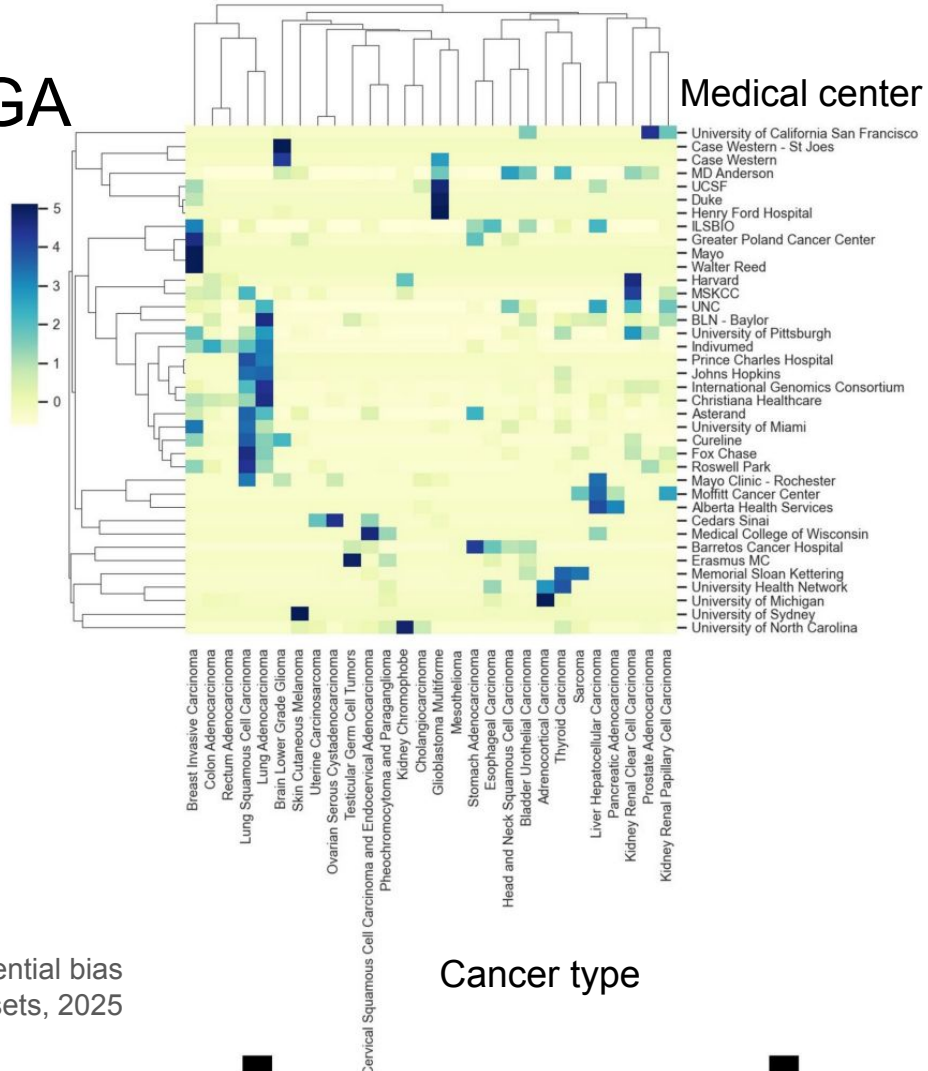
Deep Learning Can Predict Age, Scanner Type, Preparation Date, Site



Source: Schmitt, Hidden Variables in Deep Learning Digital Pathology and Their Potential to Cause Batch Effects: Prediction Model Study, 2021

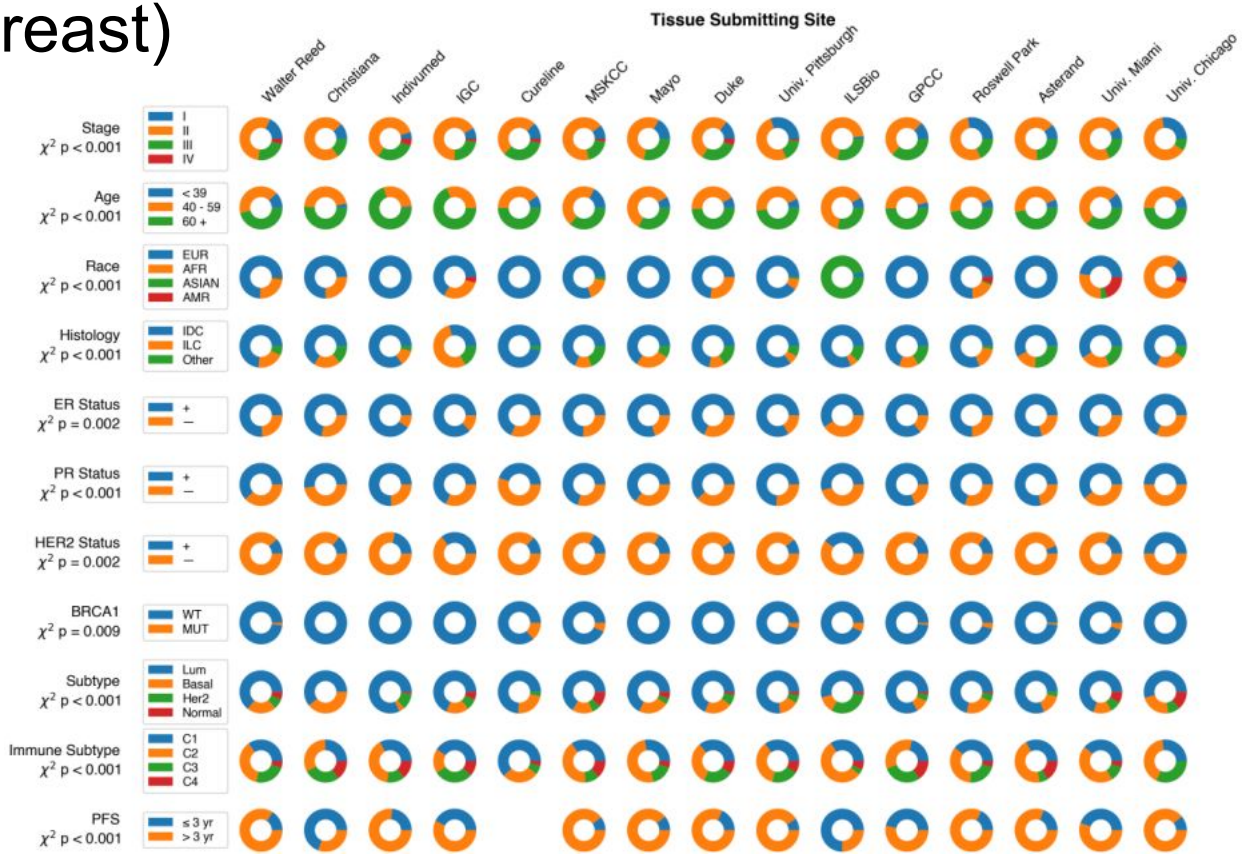
Cancer-Type Imbalance in TCGA

Prevalence of cancer types in different medical centers



Source: Kheiri, Investigation on potential bias factors in histopathology datasets, 2025

Demographic and Tumor Variations Across Sites (TCGA Breast)

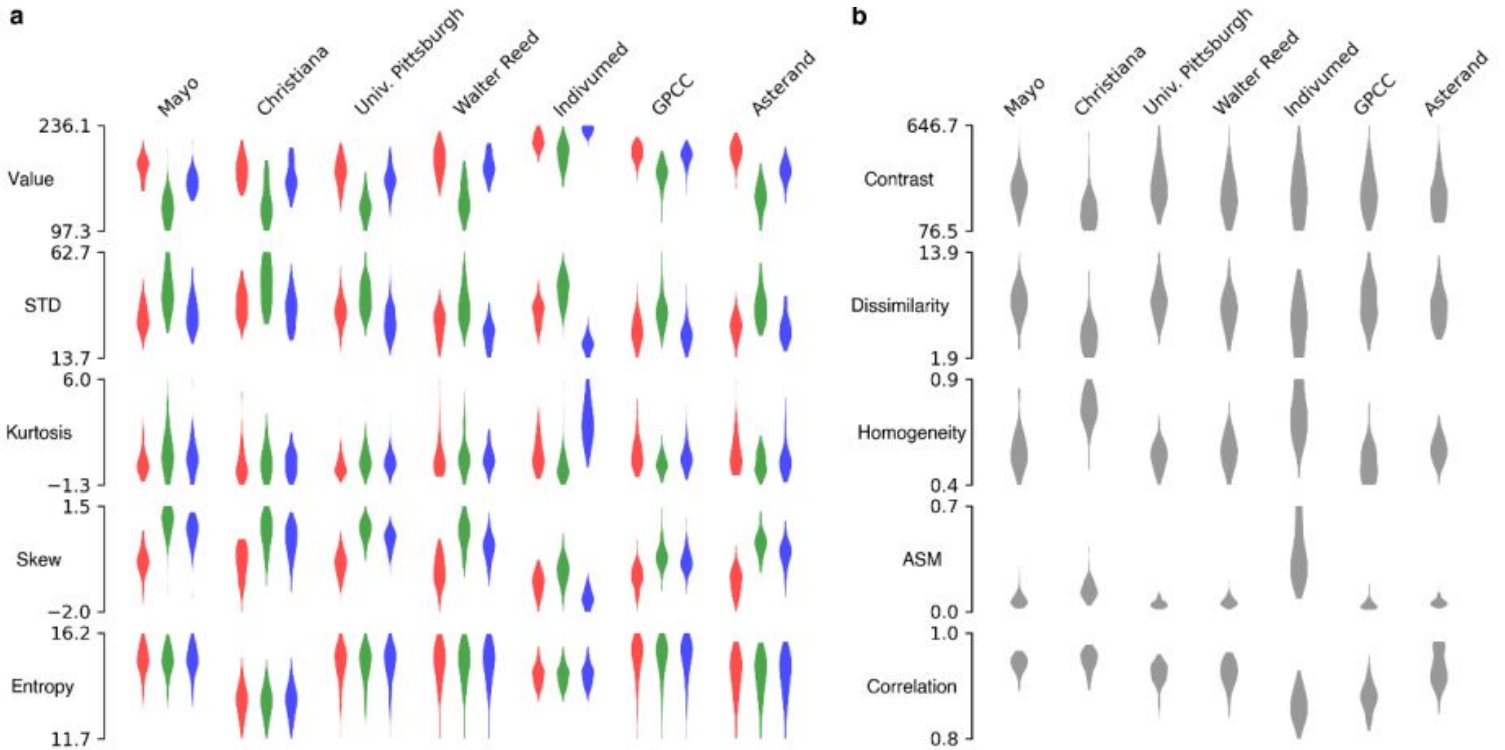


Source: Howard, The Impact of Digital Histopathology Batch Effect on Deep Learning Model Accuracy and Bias, 2020

Color Variability in TCGA

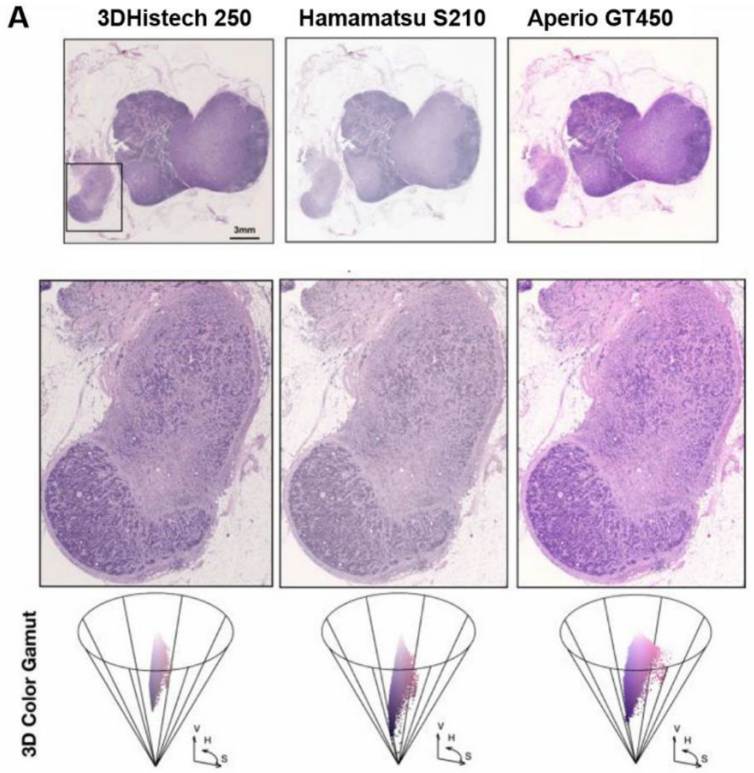
RGB statistics

Haralick texture features



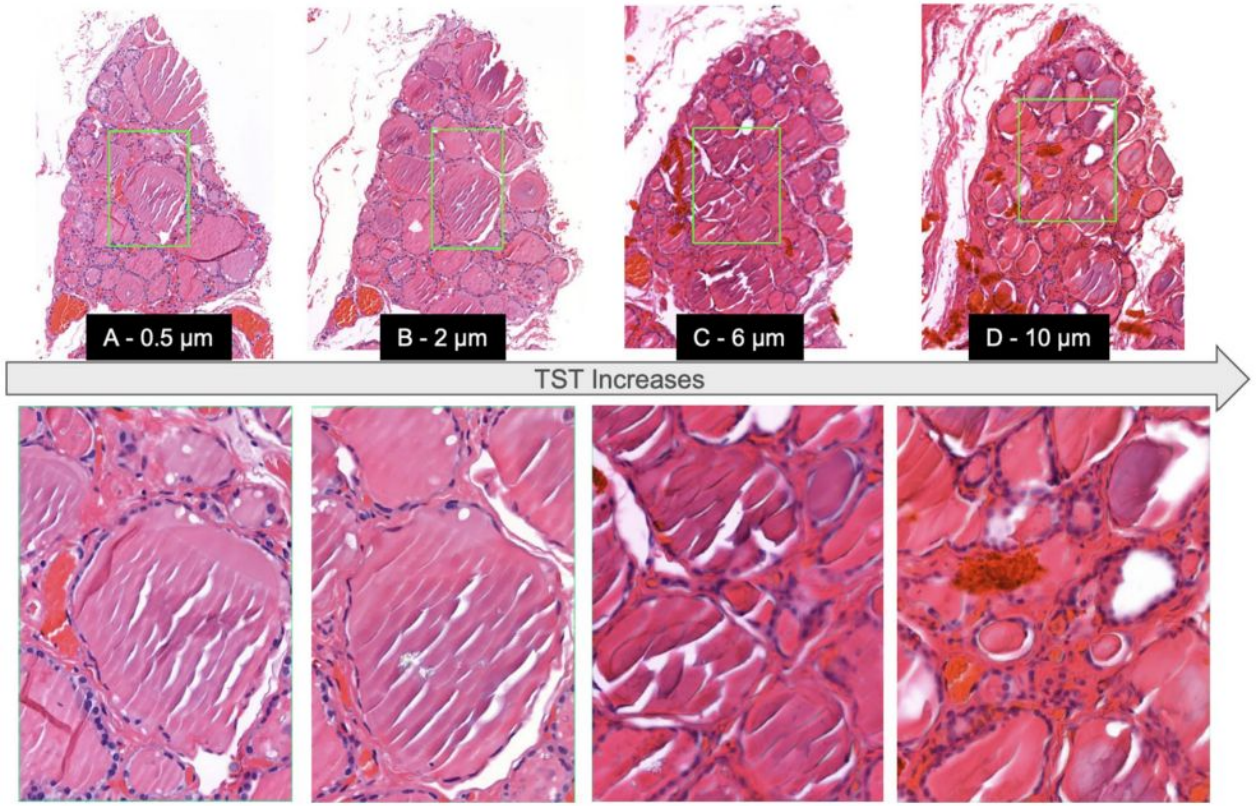
Source: Howard, The Impact of Digital Histopathology Batch Effect on Deep Learning Model Accuracy and Bias, 2020

Scanner Variability



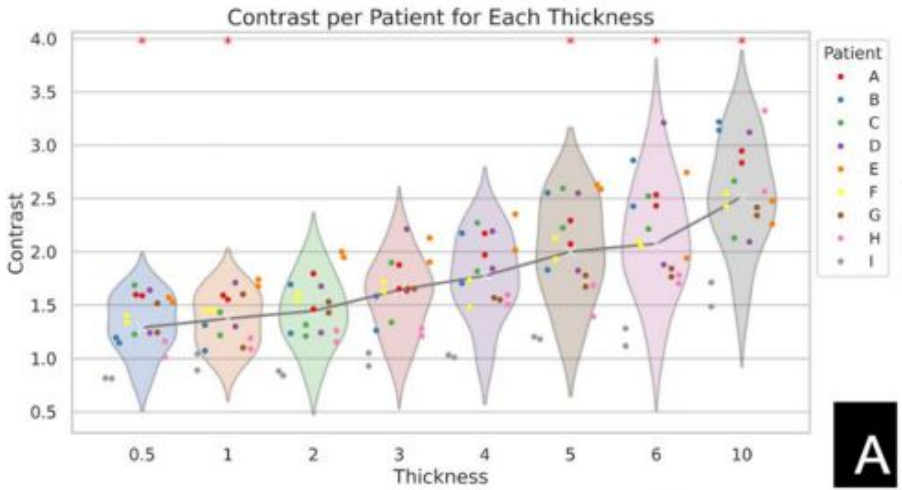
Source: Chen, Algorithm fairness in artificial intelligence for medicine and healthcare, 2023

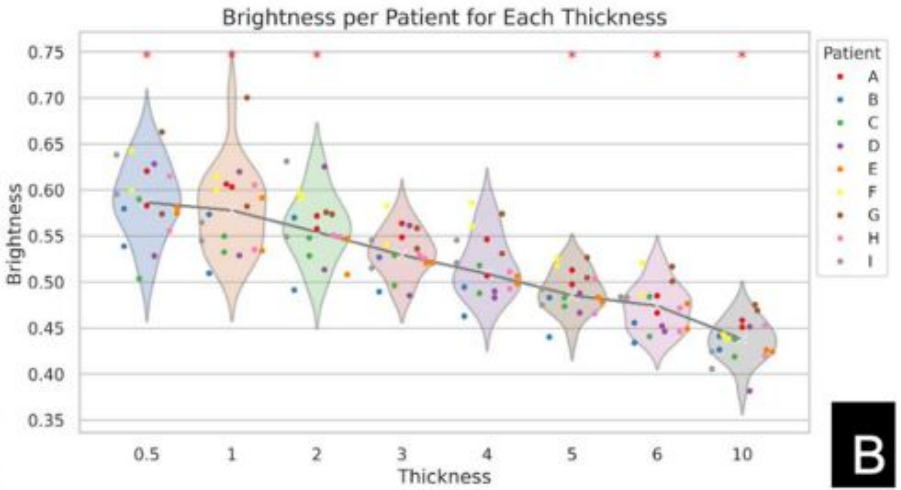
Tissue Thickness Variations



Source: Shah, Impact of Tissue Thickness on Computational Quantification of Features in Whole Slide Images for Diagnostic Pathology, 2025

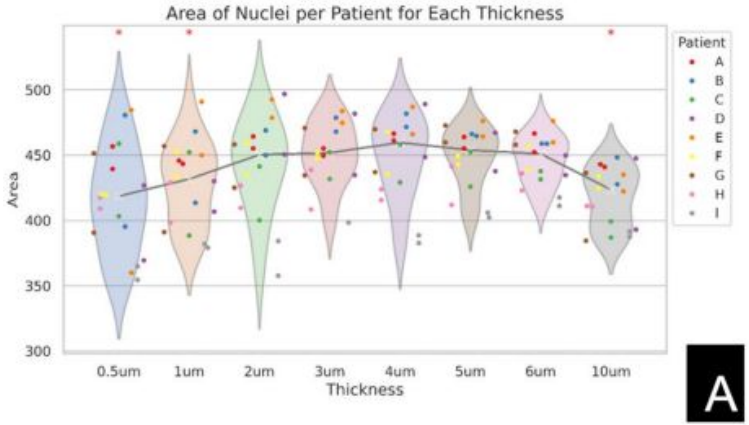
Tissue Thickness Effects Color

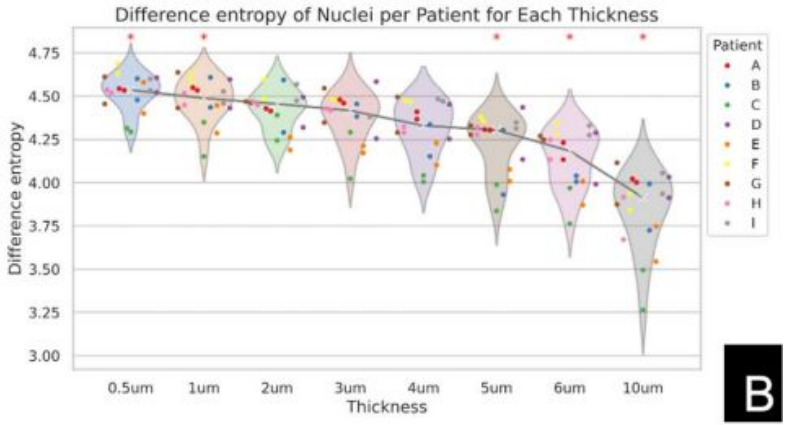


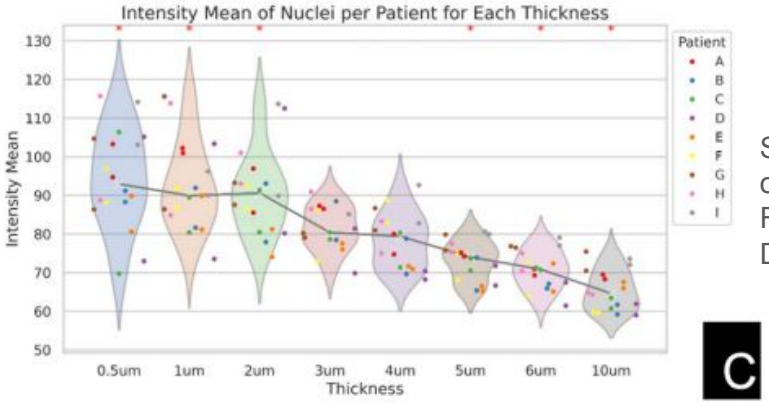


Source: Shah, Impact of Tissue Thickness on Computational Quantification of Features in Whole Slide Images for Diagnostic Pathology, 2025

Tissue Thickness Effects Nuclei

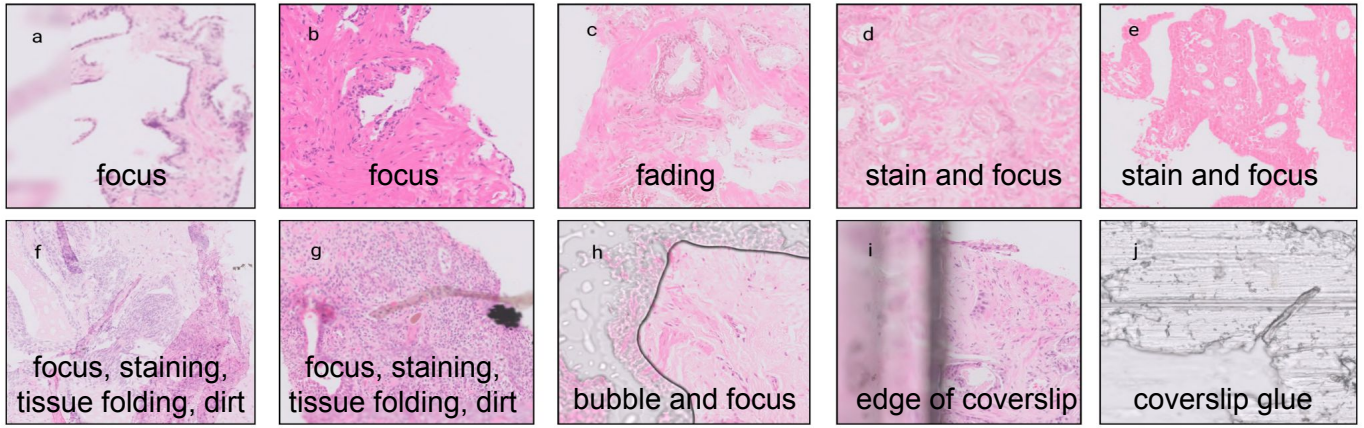


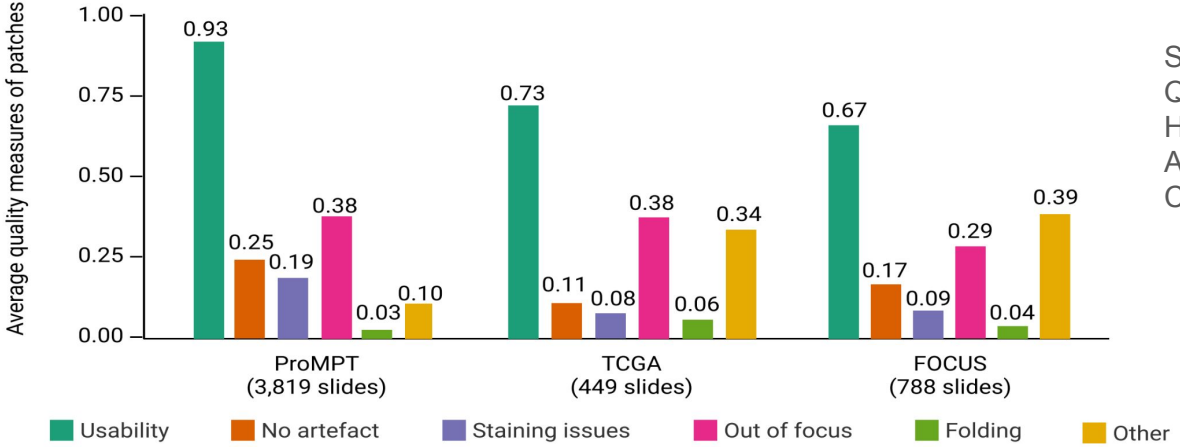




Source: Shah, Impact of Tissue Thickness on Computational Quantification of Features in Whole Slide Images for Diagnostic Pathology, 2025

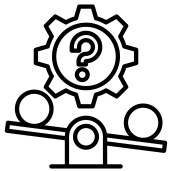
Slide Preparation and Image Acquisition Artifacts





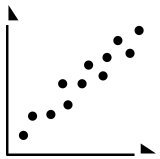
Source: Haghighat, PathProfiler: Automated Quality Assessment of Retrospective Histopathology Whole-Slide Image Cohorts by Artificial Intelligence – A Case Study for Prostate Cancer Research, 2021

Terminology

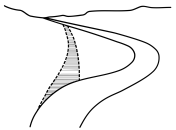


Bias: systematic deviation from fairness or accuracy that leads to partial or prejudiced outcomes

Batch effect: systematic difference in data caused by technical variations between groups of samples processed at different times or places, rather than by true biological differences



Spurious correlation: statistical relationship between two variables that appears significant but is actually caused by a third confounding variable rather than any direct causal link between the variables

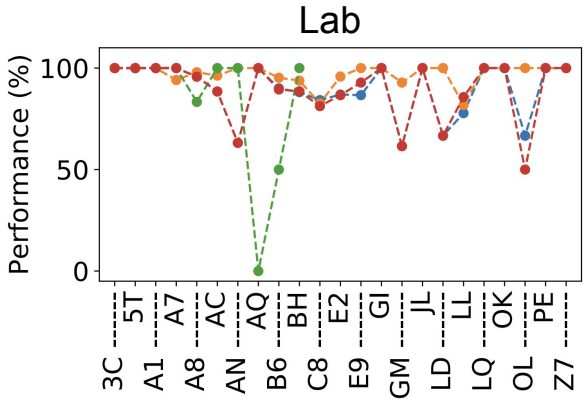


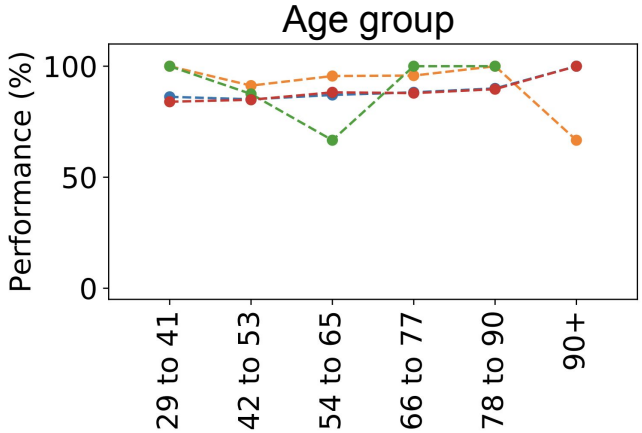
Shortcut: model relies on simple, superficial correlations to make predictions rather than learning the true underlying relationships relevant to the task

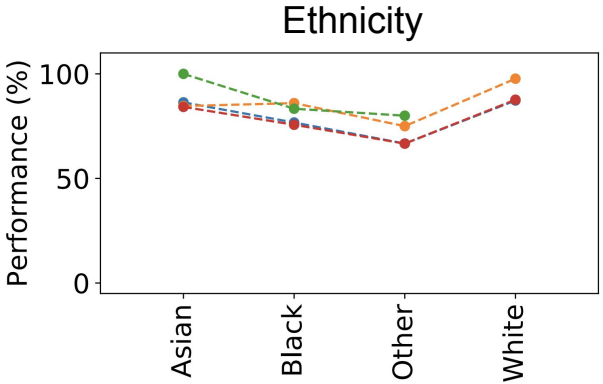
How to Detect Bias, Batch Effects, and Spurious Correlations?

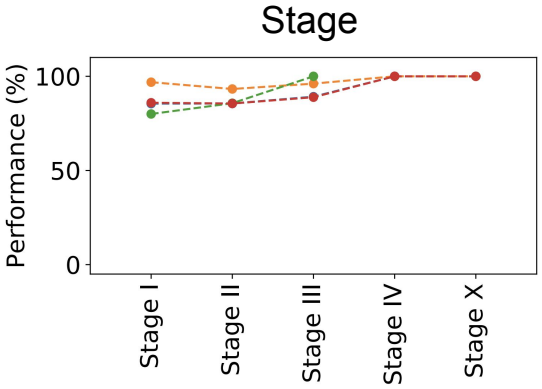
- Talk to domain experts
- Exploratory data analysis:
  - Data imbalance
  - Image variability
  - Batch effects
- Calculate stratified metrics
- Test for spurious correlations

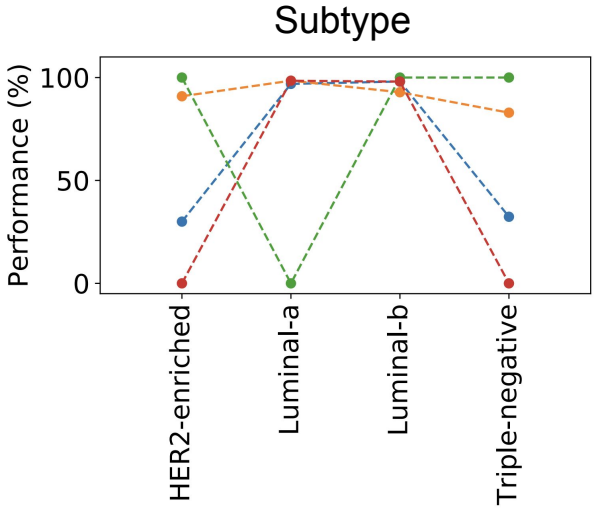
Stratify Metrics to Check for Bias









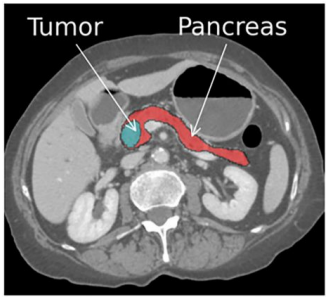


Source: Arslan, Evaluation of a predictive method for the H&E-based molecular profiling of breast cancer with deep learning, 2022

Detecting Spurious Correlations

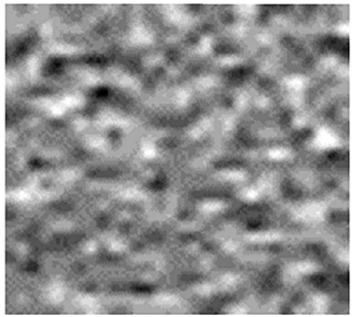
Sanity test	Implications of failing the test
Train and test with and without the target: The system should achieve an AUC of around 0.5 when tested without the target in test images.	Images contain spurious covariates that can be exploited by the model.
Train and test using noise images: The system should achieve an AUC of around 0.5 on test data.	Classification performance cannot be attributed to recognition of the target (i.e., covariates contribute to the learned classification decision rule).
Test system with different sized ROIs: The additional or reduced context should not alter the performance.	The system cannot decorrelate features of the target from its co-occurring context [i.e., Contextual Bias (55)].

Original with pancreas (WP)

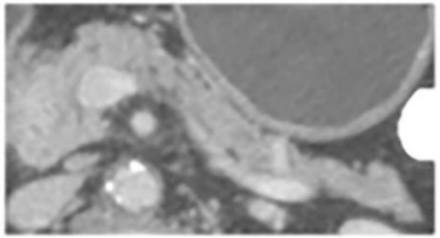


Original without pancreas (WOP)

Noise image (generated from slice differences)



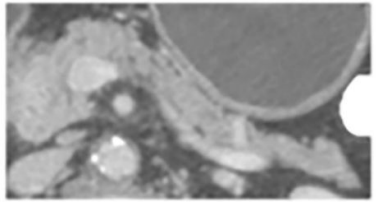
Pancreas only



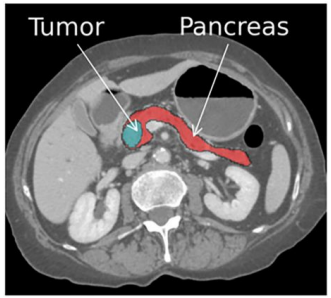
Source: Mahmood, Detecting Spurious Correlations With Sanity Tests for Artificial Intelligence Guided Radiology Systems, 2021

Detecting Spurious Correlations

Pancreas only

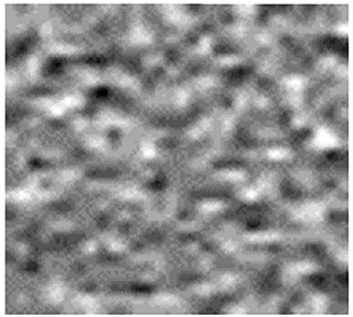


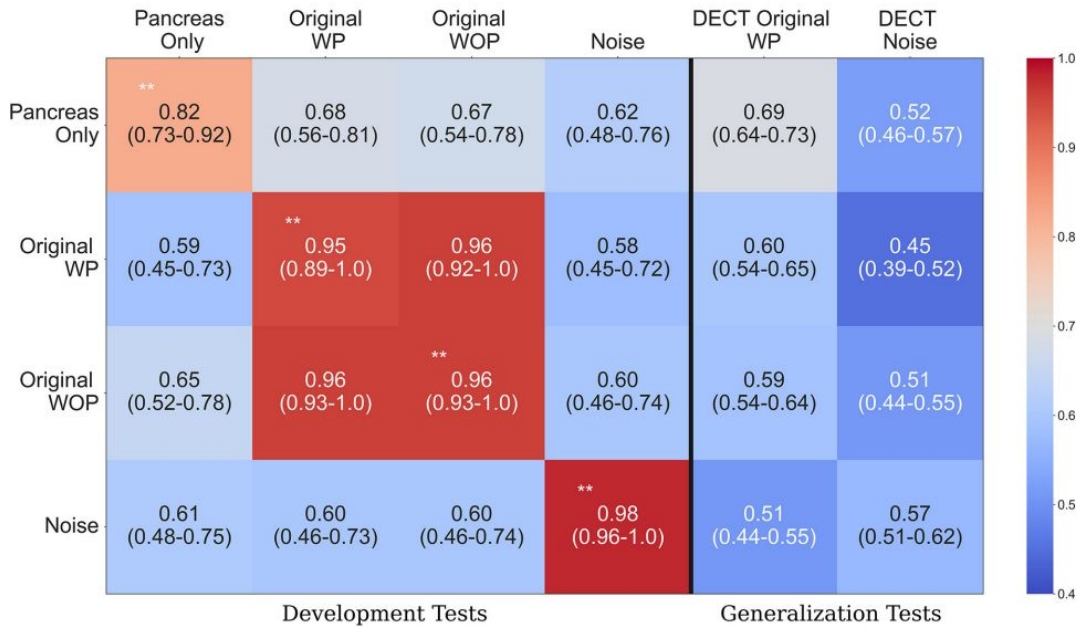
Original with pancreas (WP)



Original without pancreas (WOP)

Noise image (generated from slice differences)





Source: Mahmood, Detecting Spurious Correlations With Sanity Tests for Artificial Intelligence Guided Radiology Systems, 2021

Mitigation

Validation-level interventions: careful cross-validation

Data-level interventions: better collection, preprocessing, augmentation (modality-dependent)

Model-level interventions: strategic sampling, group-specific model, feature normalization, architecture choices, adversarial learning, foundation models

Careful Cross-Validation

Source: Howard, The Impact of Digital Histopathology Batch Effect on Deep Learning Model Accuracy and Bias, 2020

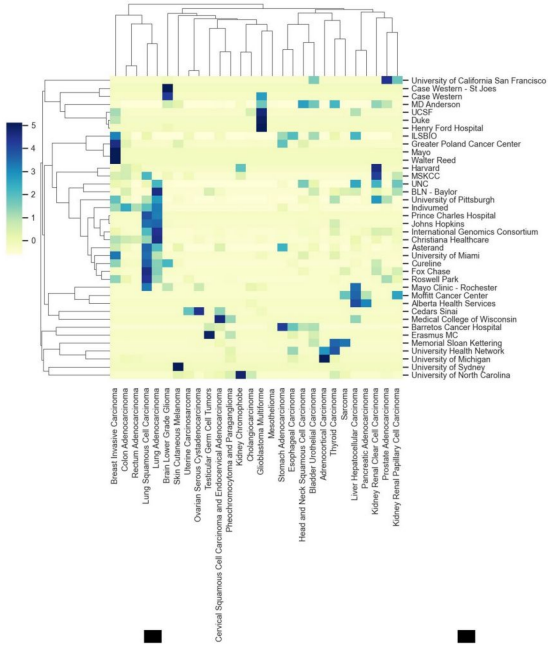
Strategic Sampling or Subtype-Specific Model

Homologous Recombination Deficiency
in luminal breast cancers

Dataset	Method	AUC
TCGA	No correction Strategic sampling	0.71 0.63
Curie	No correction Luminals only	0.88 0.83

Source: Lazard, Deep learning identifies morphological patterns of homologous recombination deficiency in luminal breast cancers from whole slide images, 2022

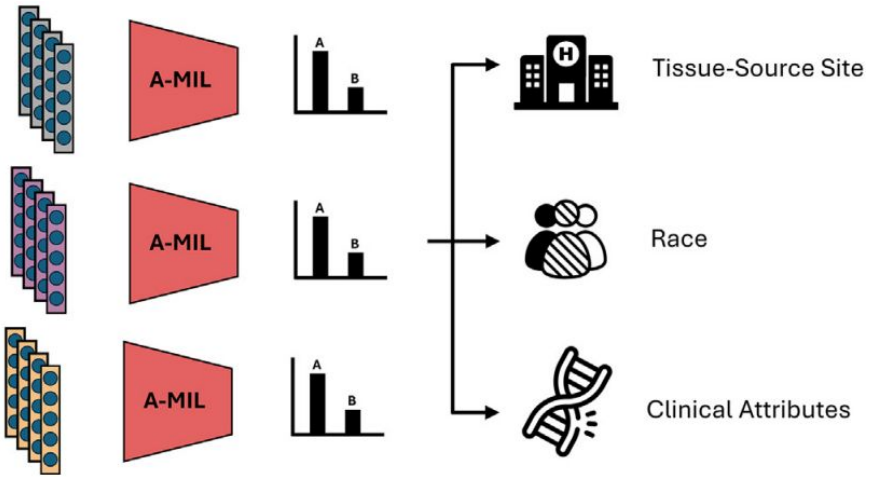
Balanced Sampling



	Cancer_num	Center_num	Cancer-Acc (%)	Center-Acc (%)
Fair subset	2	2	79	66
Correlated subset	2	2	95	95

Source: Kheiri, Investigation on potential bias factors in histopathology datasets, 2025

Batch Normalization



ComBat

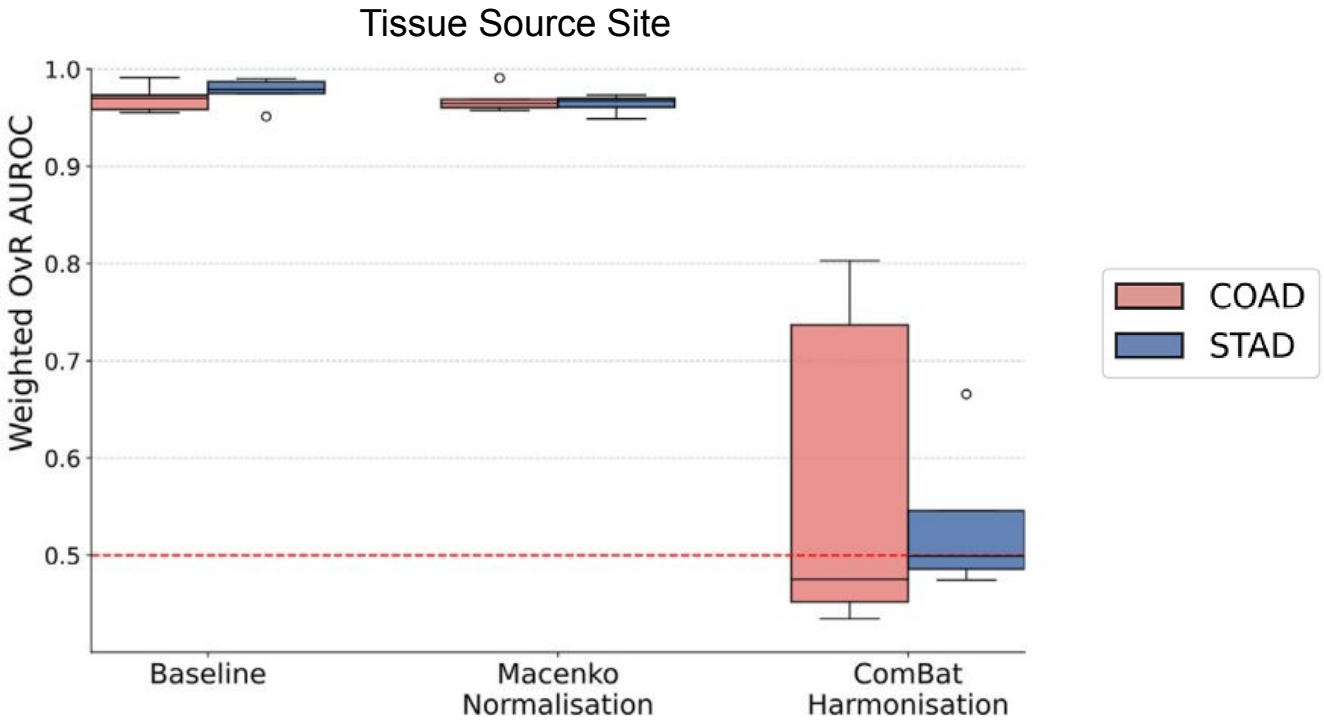
$$Y_{ijg} = \alpha_g + \gamma_{ig} + \delta_{ig}\varepsilon_{ijg}$$

$$Y^*_{ijg} = \frac{Y_{ijg} - \hat{\alpha}_g - \hat{\gamma}_{ig}}{\hat{\delta}_{ig}} + \hat{\alpha}_g$$

Feature g, Sample j, Batch i

Source: Murchan, Deep feature batch correction using ComBat for machine learning applications in computational pathology, 2024

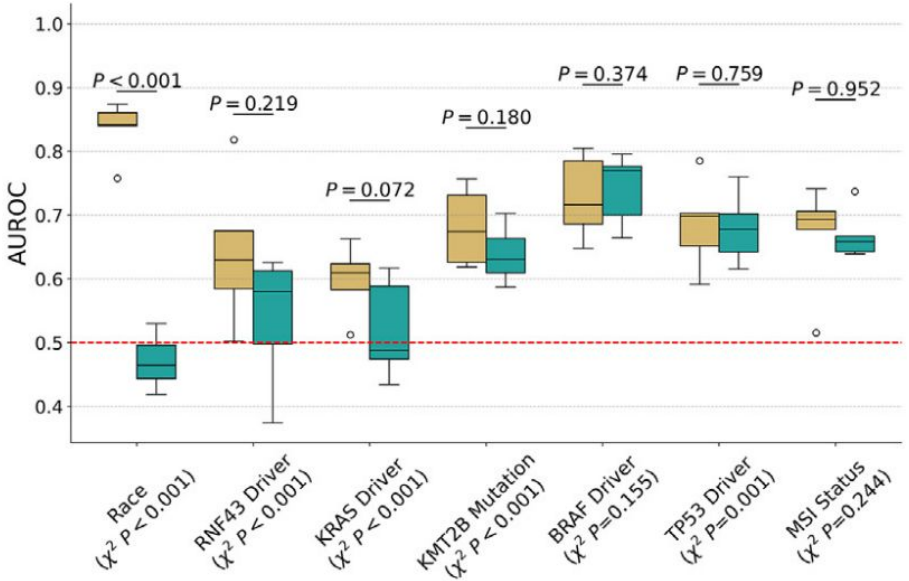
Batch Normalization



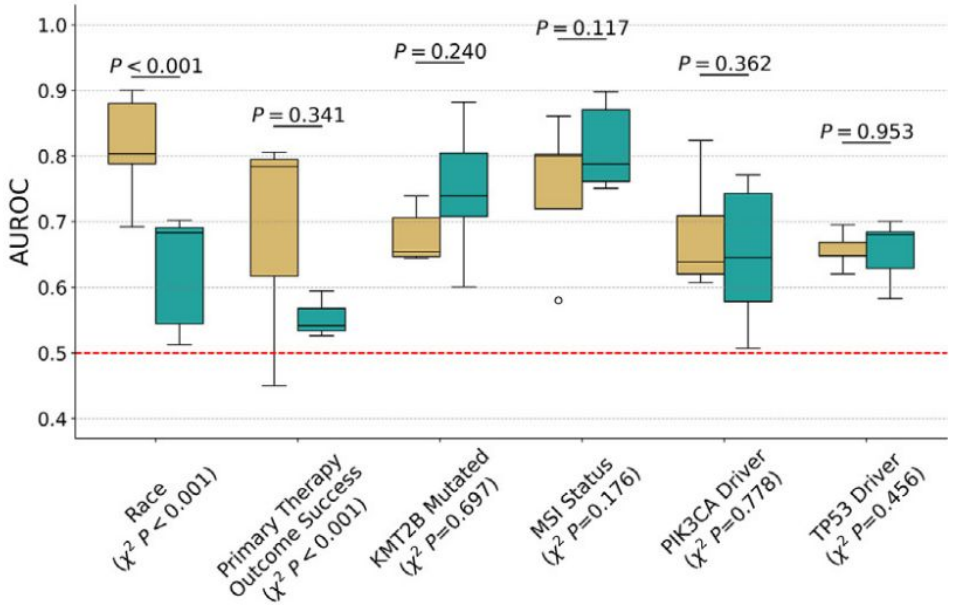
Source: Murchan, Deep feature batch correction using ComBat for machine learning applications in computational pathology, 2024

Batch Normalization

Colon Adenocarcinoma

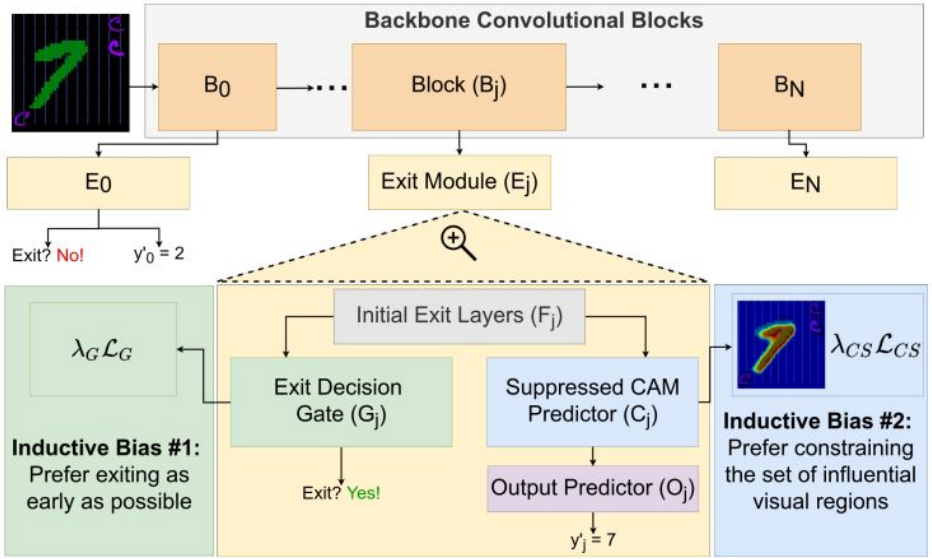


Stomach Adenocarcinoma



Source: Murchan, Deep feature batch correction using ComBat for machine learning applications in computational pathology, 2024

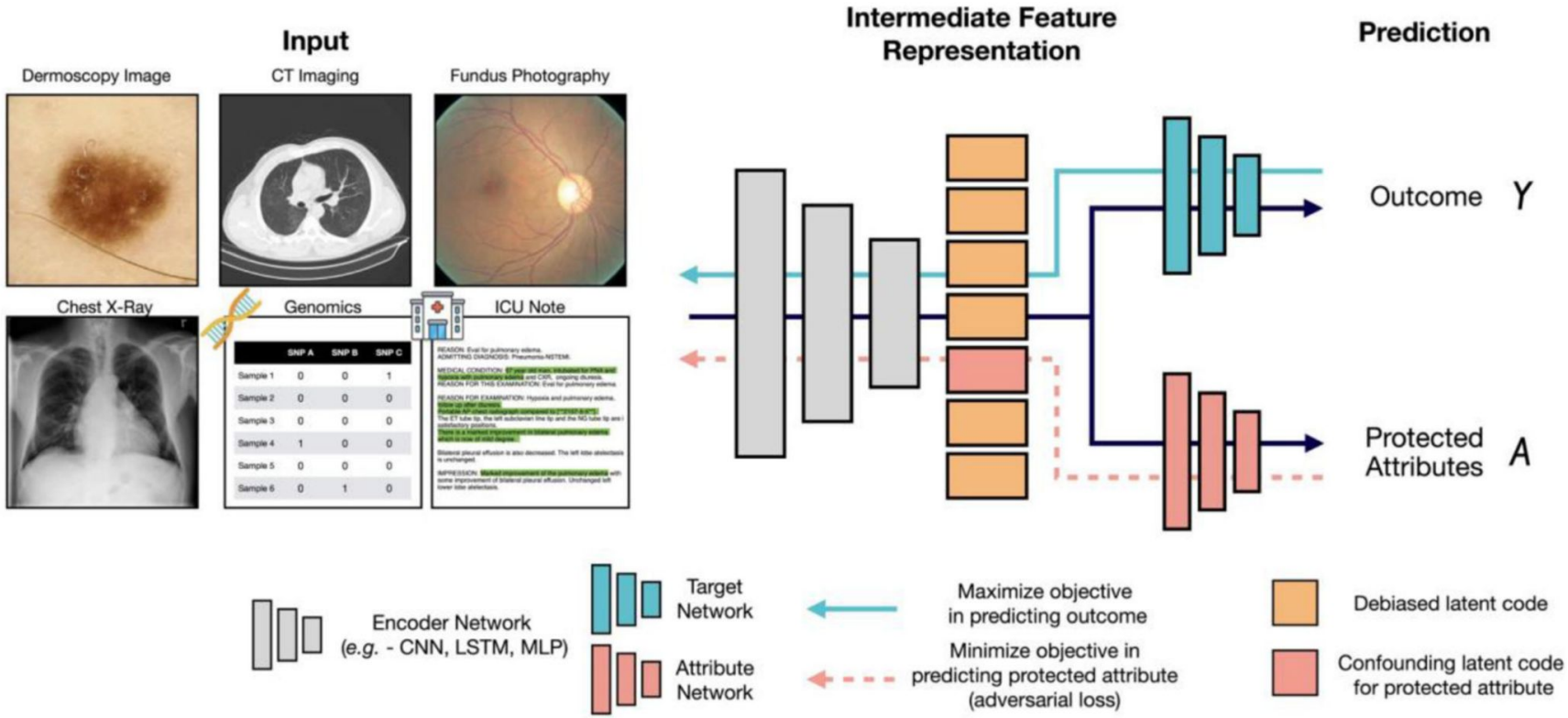
Favor Simpler Solutions



Architecture+Method	Biased MNISTv2	COCO-on-Places	BAR
Results on Standard ResNet-18			
ResNet+ERM	36.8 ± 0.7	35.6 ± 1.0	51.3 ±1.9
ResNet+SD [51]	37.1 ± 1.0	35.4 ± 0.5	51.3 ±2.3
ResNet+Up Wt	37.7 ± 1.6	35.2 ± 0.4	51.1 ±1.9
ResNet+gDRO [56]	19.2 ± 0.9	35.3 ±0.1	38.7 ±2.2
ResNet+PGI [3]	**48.6** ± 0.7	**42.7** ± 0.6	**53.6** ±0.9
Results on OccamResNet-18			
OccamResNet	**65.0** ±1.0	**43.4** ± 1.0	**52.6** ±1.9

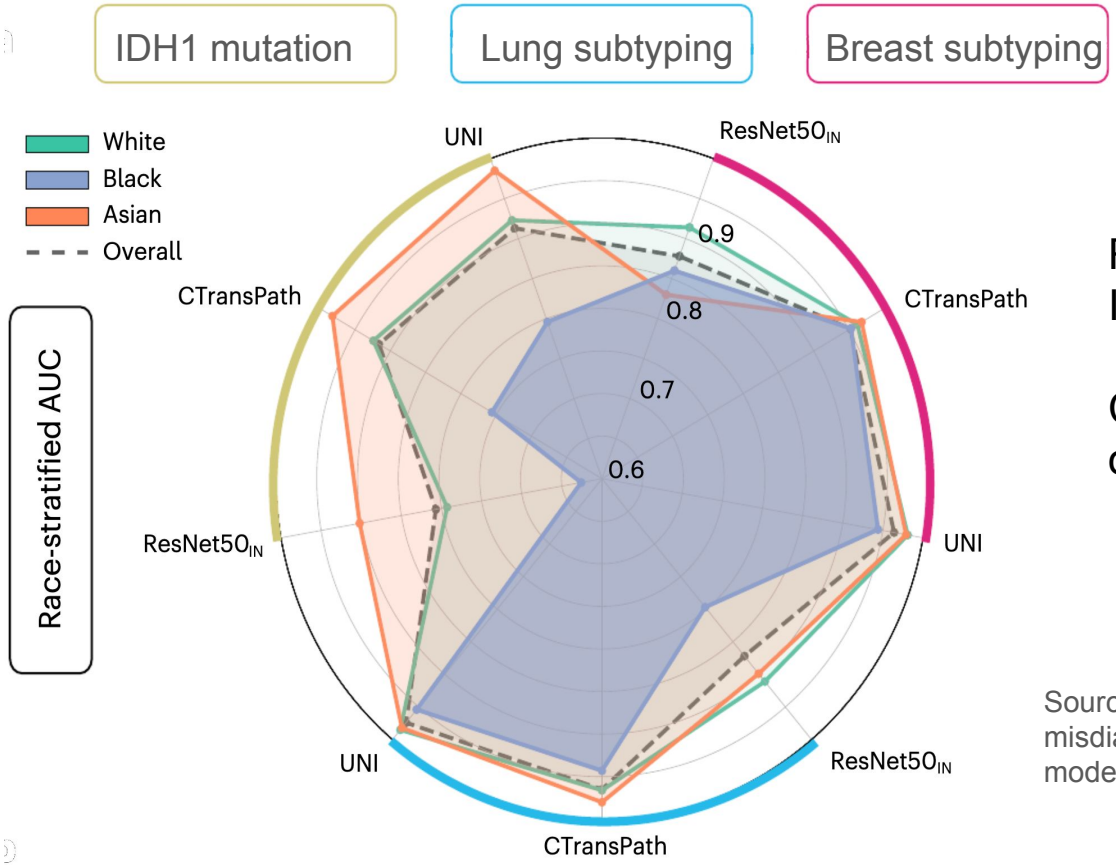
Source: Shrestha, OccamNets: Mitigating Dataset Bias by Favoring Simpler Hypotheses, 2024

Adversarial Learning



Source: Chen, Algorithm fairness in artificial intelligence for medicine and healthcare, 2023

Foundation Models Can Help Reduce Bias

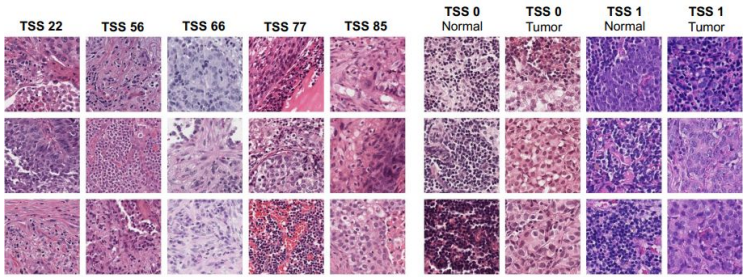


ResNet50$_{IN}$ pretrained on ImageNet

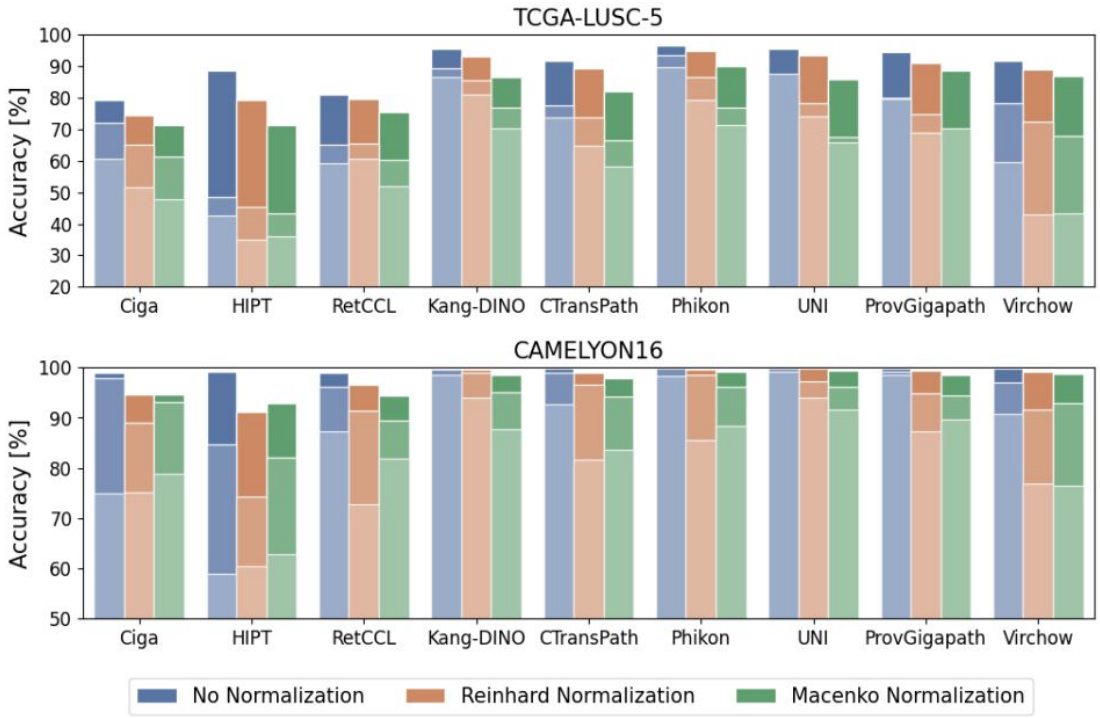
CTransPath/UNI pretrained on histopathology

Source: Vaidya, Demographic bias in misdiagnosis by computational pathology models, 2024

Foundation Models Still Encode Batch Effects



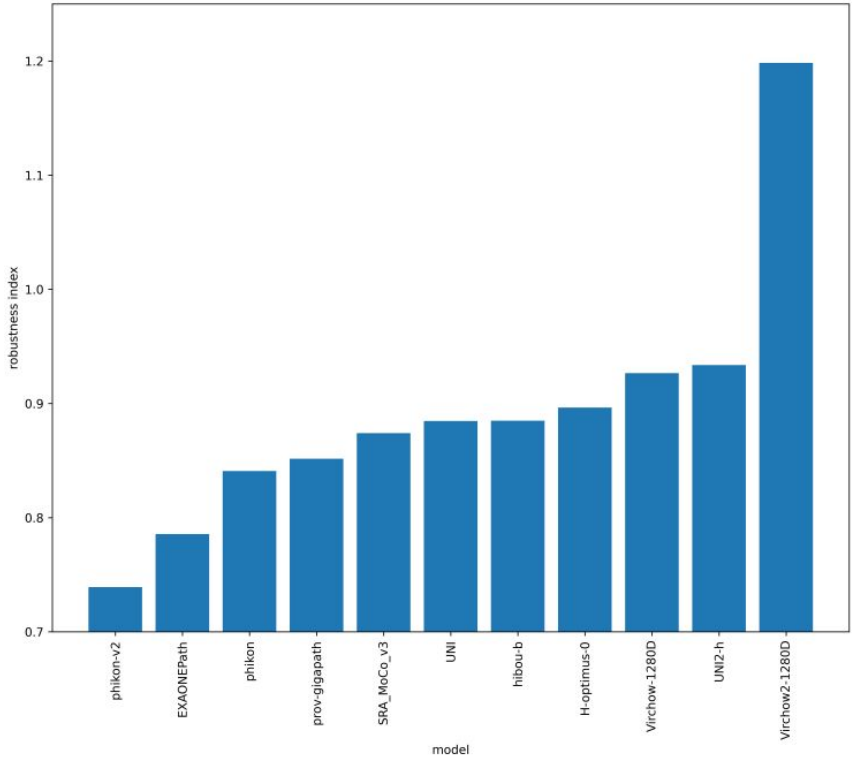
Tissue source site prediction



Source: Kömen, Do Histopathological Foundation Models Eliminate Batch Effects? A Comparative Study, 2024

Some Models are More Robust to Medical Center Differences

medical center robustness index = $\frac{\text{how many of the k nearest neighbors represent the same biological class}}{\text{how many of the k nearest neighbors represent the same medical center}}$



Source: de Jong, Current Pathology Foundation Models are Unrobust to Medical Center Differences, 2025

Key Takeaways

1. Understand your data and sources of bias: quantify data imbalance, batch effects, and potential spurious correlations early
2. Validate carefully
3. There are a variety of data- and model-level solutions, but no one-size-fits-all
4. Foundation models may help but do not solve the bias problem

Bonus Offer: 1 Hour Strategy Session

- Get unstuck with a clear set of next steps
- Improve the accuracy of your model
- Train a more robust and generalizable model
- Apply best practices for your unique challenges
- Get computer vision insights from an experienced research scientist

$500 ~~$750~~ (33% discount if booked in the next 30 days)

https://calendly.com/hdcouture/post-webinar-strategy-session

Other inquiries: heather@pixelscientia.com

Pixel Scientia LABS