AIM 2: Artificial Intelligence in Medicine II

Harvard - BMI 702 and BMIF 203, Spring 2026
Lecture 6: Medical Imaging II

Vision foundation models. Development and validation of medical imaging interpretation models, Model robustness and performance across diverse populations.

HARVARD
MEDICAL SCHOOL

Lecture Outline

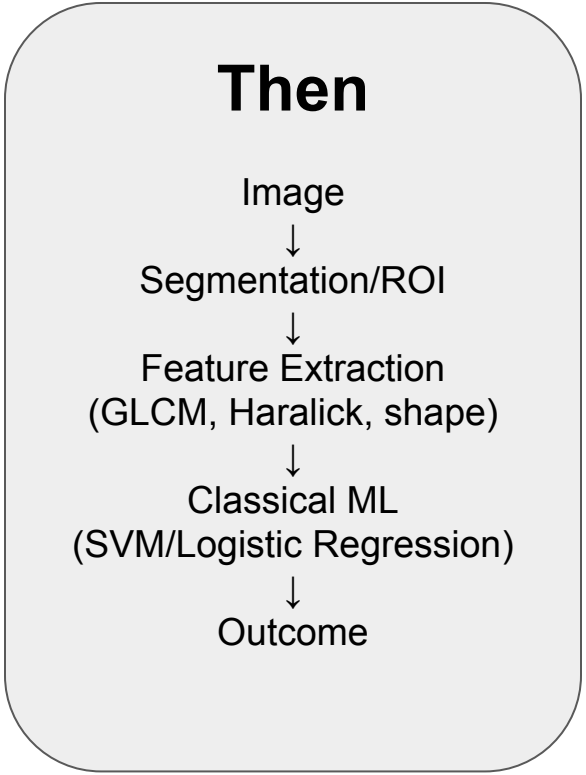
1. Foundation Models
2. Pathology Foundational Models
3. Model Interpretability
4. Model Generalizability

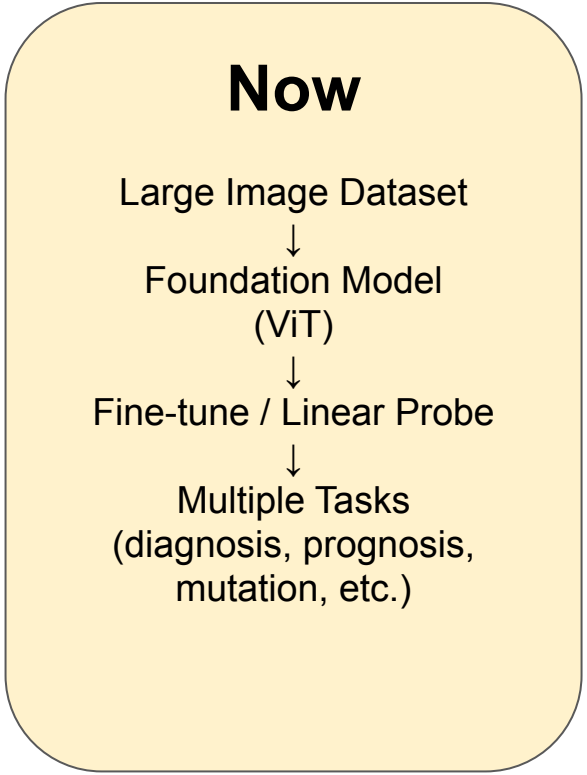
Lecture Outline

1. Foundation Models
2. Pathology Foundational Models
3. Model Interpretability
4. Model Generalizability

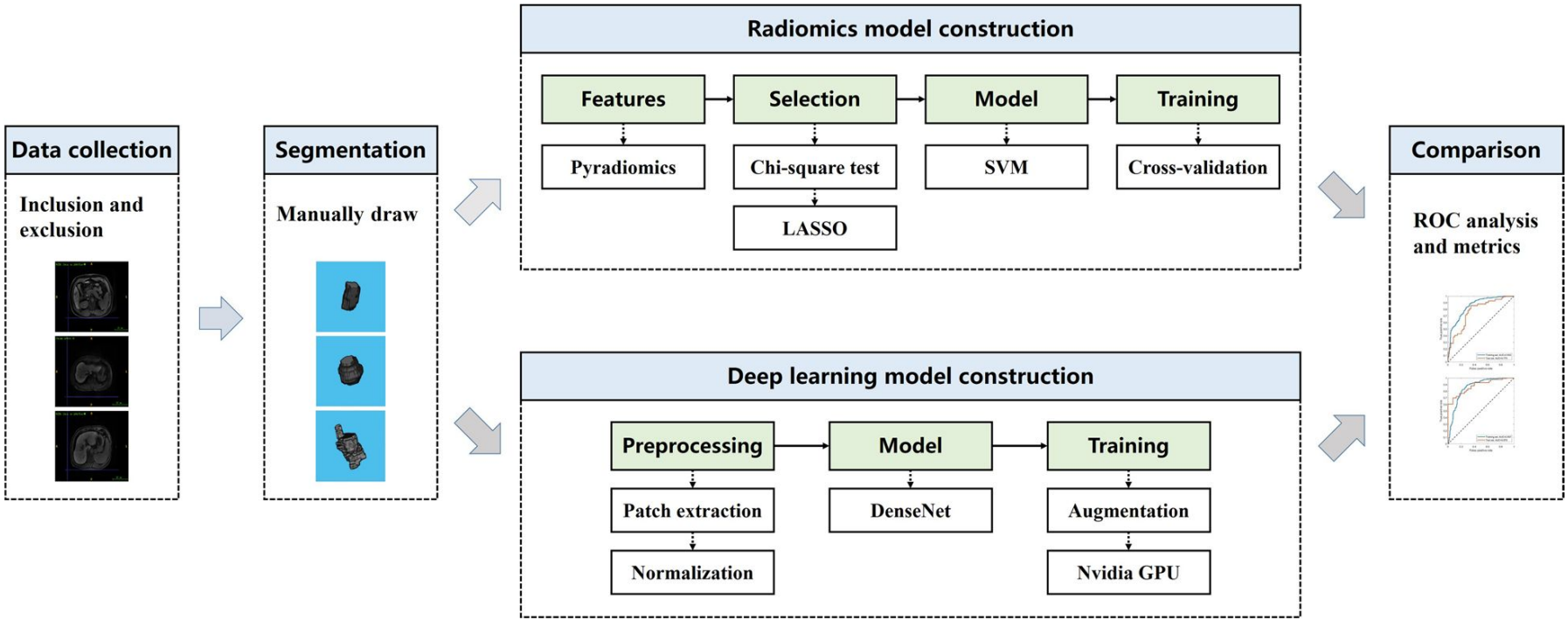
From Engineered Features to Foundation Models

- Then:
  engineered features + small supervised models
- Now:
  large-scale pretraining + adaptation
- Constant:
  data limits, domain shift, clinical reality



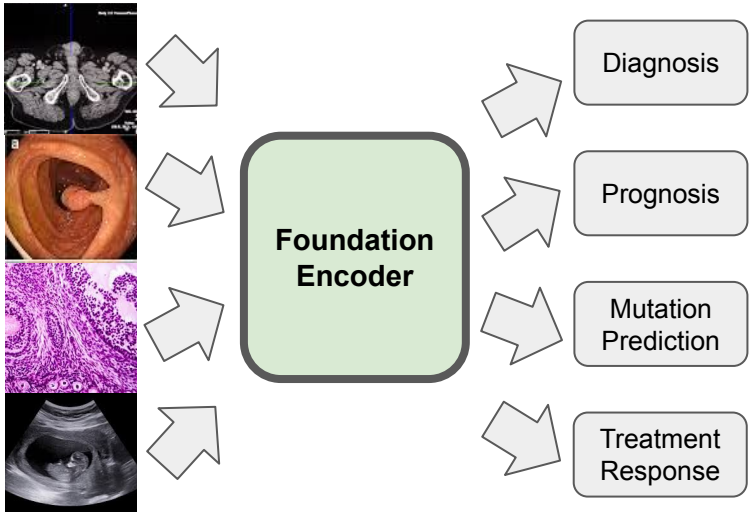


Radiomics vs Learned Representations



Du, L., Yuan, J., Gan, M., Li, Z., Wang, P., Hou, Z., & Wang, C. (2022). A comparative study between deep learning and radiomics models in grading liver tumors using hepatobiliary phase contrast-enhanced MR images. BMC Medical Imaging, 22(1), 218.

The Promise



One encoder | Many tasks | Less task-specific data

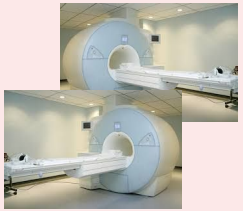
The Problem

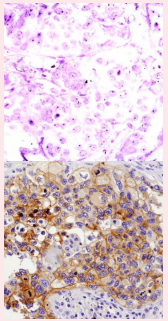
Different scanners

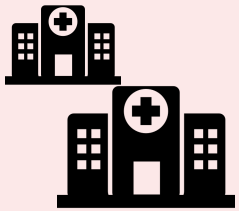
Different hospitals

Different stain colors

Slightly shifted distributions

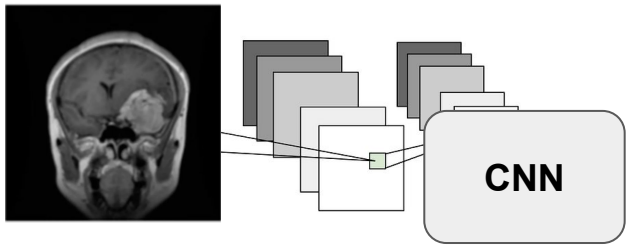


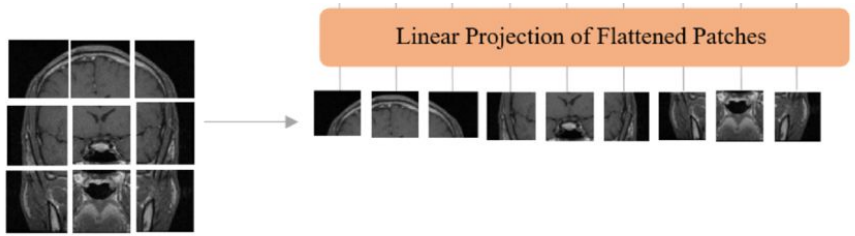




Heterogeneous modalities | Noisy labels
Site artifacts & shortcuts | Calibration & reliability

From Convolutions to Tokens





CNNs

- Local receptive fields
- Translation equivariance

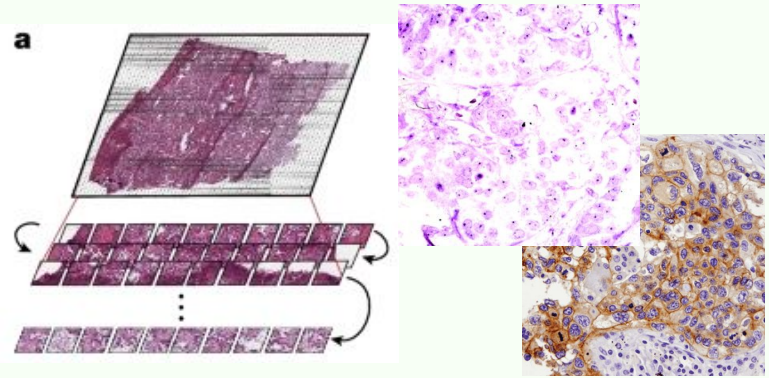
ViT

- Patchify image
- Global self-attention

Shift: *local filters → global token interactions*

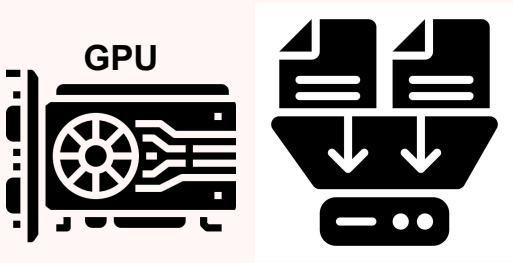
Tummala, S., Kadry, S., Bukhari, S. A. C., & Rauf, H. T. (2022). Classification of brain tumor from magnetic resonance imaging using vision transformers ensembling. *Current Oncology*, *29*(10), 7498-7511.

Patches as Tokens — Power and Cost



What it buys you

- Global context
- Flexible scale
- Modality-agnostic interface

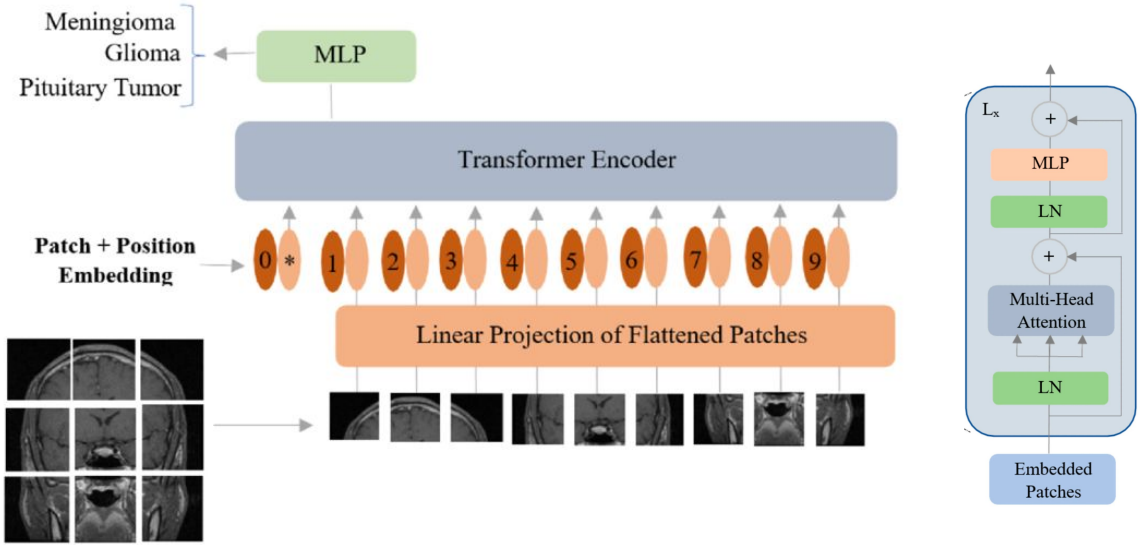


What it costs

- Data hunger
- Compute
- Sensitivity to shift

Tummala, S., Kadry, S., Bukhari, S. A. C., & Rauf, H. T. (2022). Classification of brain tumor from magnetic resonance imaging using vision transformers ensembling. *Current Oncology*, *29*(10), 7498-7511.

Inside a Vision Transformer Encoder



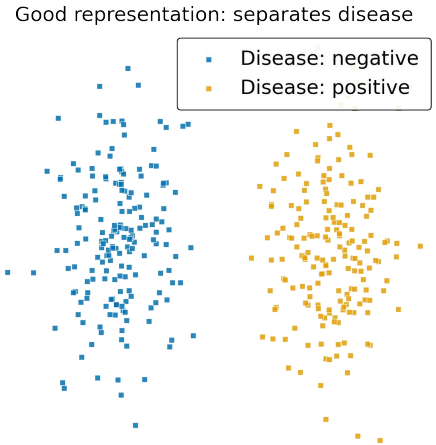
Self-Attention Mechanism

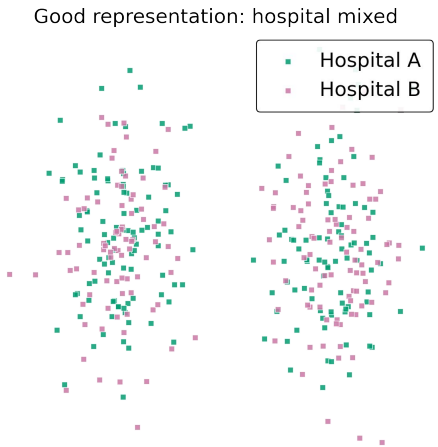
- Given Queries ($\mathbf{Q}$), Keys ($\mathbf{K}$), Values ($\mathbf{V}$) of dimension d:

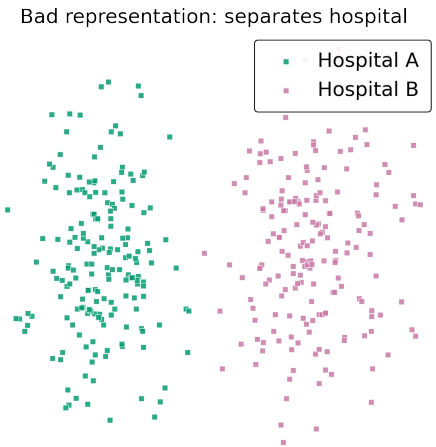
$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}}\right)\mathbf{V}$$

Tummala, S., Kadry, S., Bukhari, S. A. C., & Rauf, H. T. (2022). Classification of brain tumor from magnetic resonance imaging using vision transformers ensembling. *Current Oncology*, *29*(10), 7498-7511.

What Makes a “Good” Representation?







How Foundation Models are Built

Pretraining Signals

- Self-supervised
- Contrastive
- Weak supervision
- Multimodal alignment

Different signals → different representations

Self-Supervised Image ↓ Masked patches ↓ Reconstruction head	**Contrastive** Image A (view 1) ↔ (view 2) + Image B far away (Positive vs negative arrows)
Weak Supervision Whole slide ↓ slide label ↓ encoder	**Multimodal** Image encoder ↓ shared embedding ↑ Text encoder

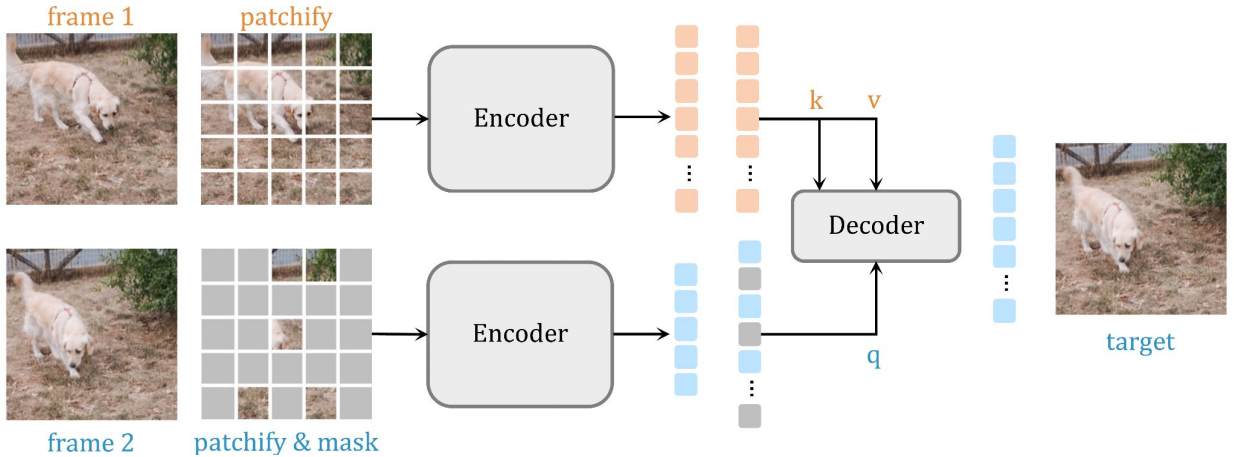
Self-Supervised Learning

Core Idea
Predict missing information from context

Examples

- MAE (Masked Autoencoders)
- BEiT
- Masked image modeling

Captures structure without labels



Gupta, A., Wu, J., Deng, J., & Li, F. F. (2023). Siamese masked autoencoders. Advances in Neural Information Processing Systems, 36, 40676-40693.

Contrastive Learning

Core Idea

Pull positives together
Push negatives apart

Frameworks

- SimCLR
- MoCo
- DINO (self-distillation)

Strong transfer performance

https://encord.com/blog/guide-to-contrastive-learning/

Weak Supervision at Scale

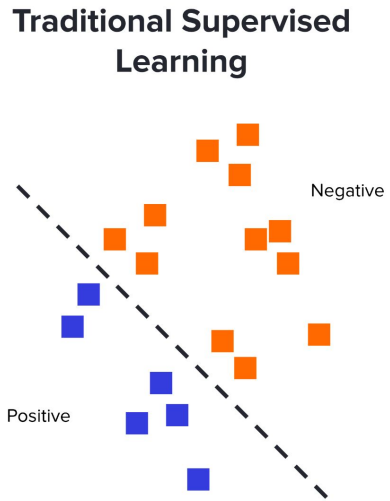
Label Sources

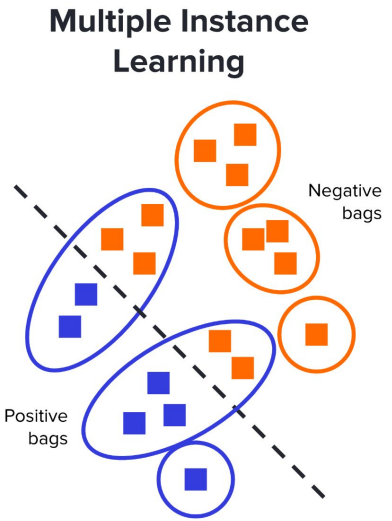
- Slide-level diagnosis
- Radiology reports
- Billing codes

Tradeoff

Scale ↑

Label quality ↓





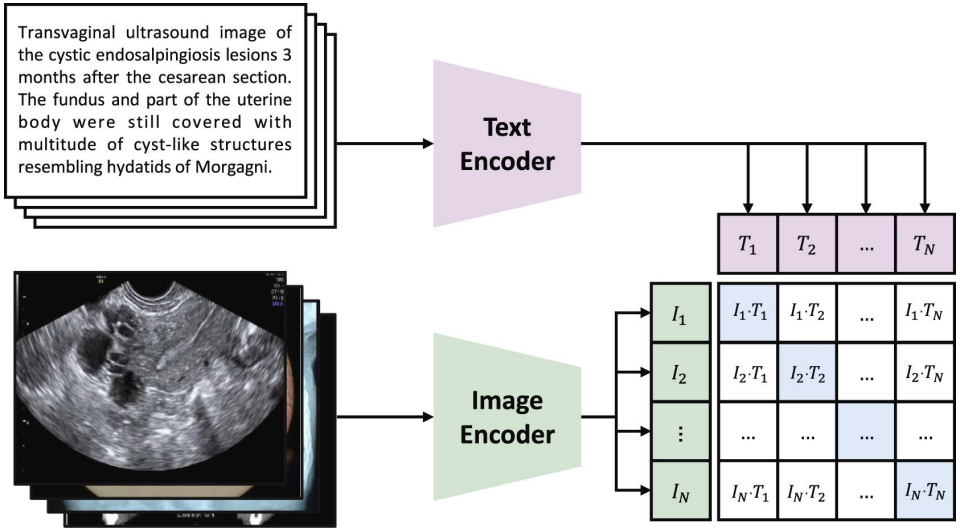
https://nilg.ai/202105/an-introduction-to-multiple-instance-learning/

Multimodal Alignment

Align images to language or omics

Image encoder ↔ Text/omics encoder

Toward semantic representations



Zhao, Z., Liu, Y., Wu, H., Wang, M., Li, Y., Wang, S., ... & Shen, D. (2025). CLIP in medical imaging: A survey. *Medical Image Analysis*, *102*, 103551.

Lecture Outline

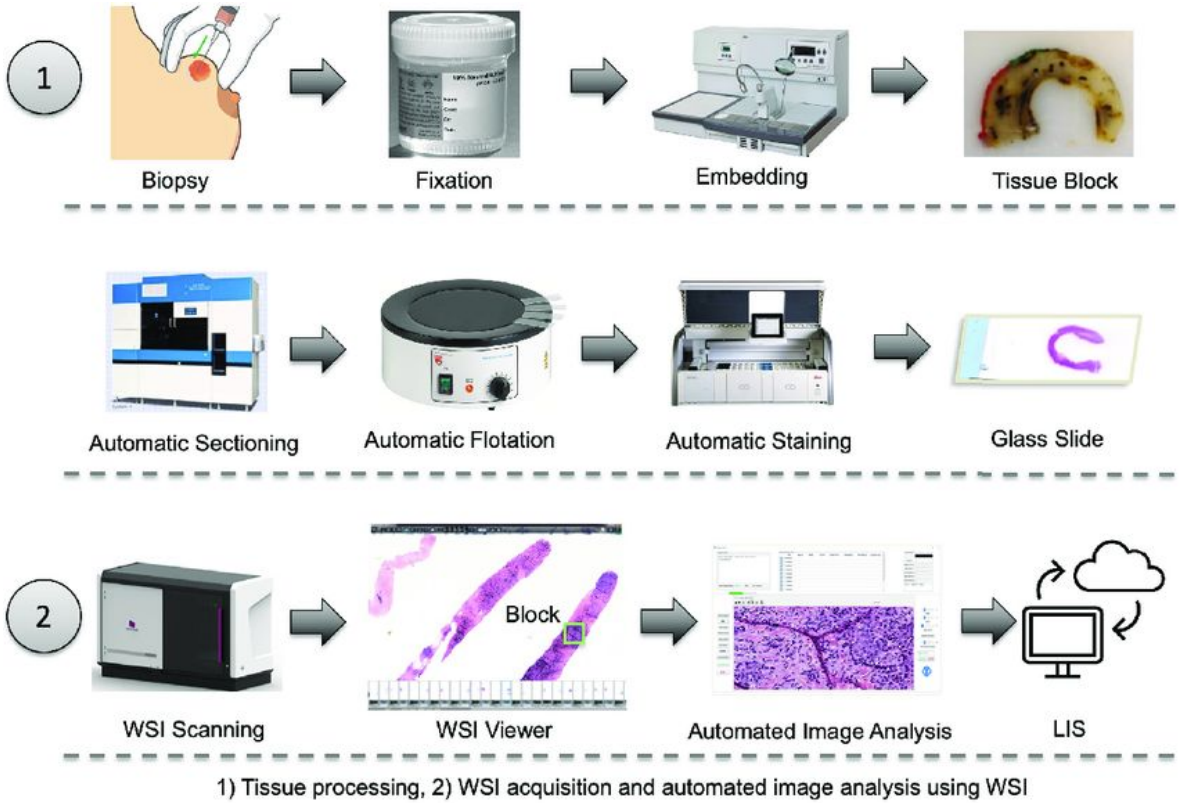
1. ~~Foundation Models~~
2. Pathology Foundational Models
3. Model Interpretability
4. Model Generalizability

Why Pathology Became the Scaling Laboratory

Digital Pathology Enables Scale

- Gigapixel whole-slide images
- Routine digitization → massive archives
- Slide-level labels available at scale

Pathology became the foundation-model sandbox



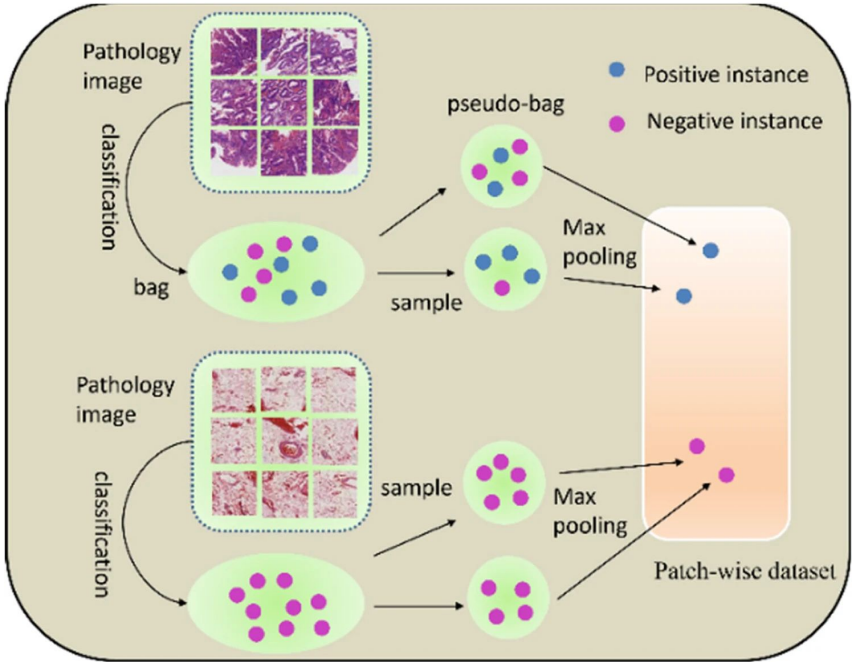
1) Tissue processing, 2) WSI acquisition and automated image analysis using WSI

Whole-Slide Images as “Bags of Patches”

WSI → Many Patches → One Label

- Supervision at slide level
- Signal is sparse
- Aggregation is the challenge

Multiple-instance learning framing



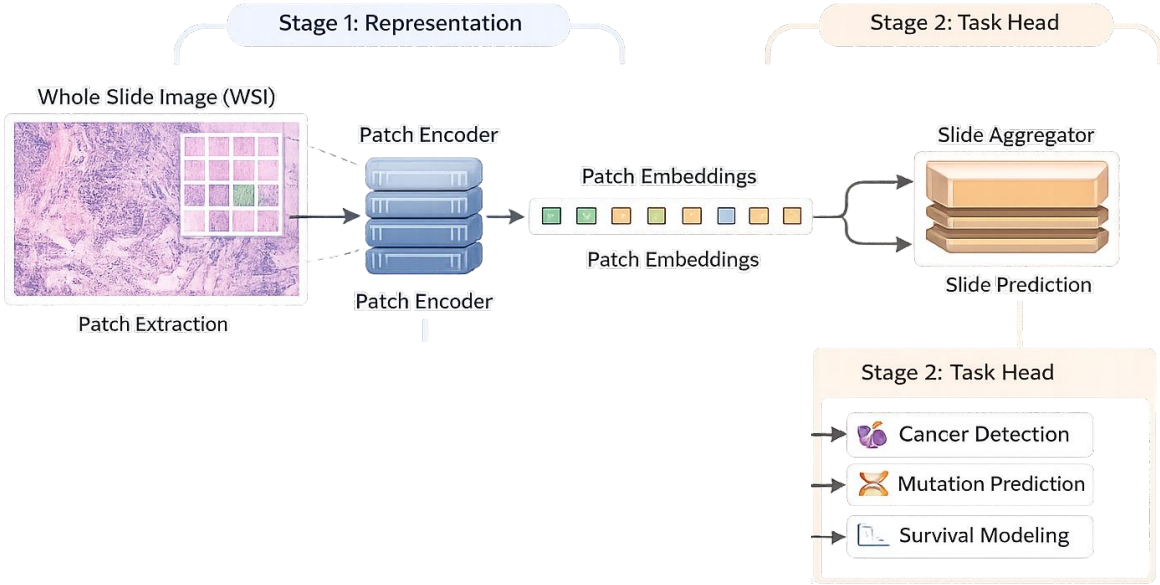
Yang, B., Ding, L., Li, J., Li, Y., Qu, G., Wang, J., ... & Liu, B. (2025). Transformer-based multiple instance learning network with 2D positional encoding for histopathology image classification. Complex & Intelligent Systems, 11(5), 218.

Patch Encoder + Slide Aggregator

Two-Stage System

1. Patch encoder (representation learning)
2. Slide aggregator (task-specific pooling)

Foundation move: make encoder reusable

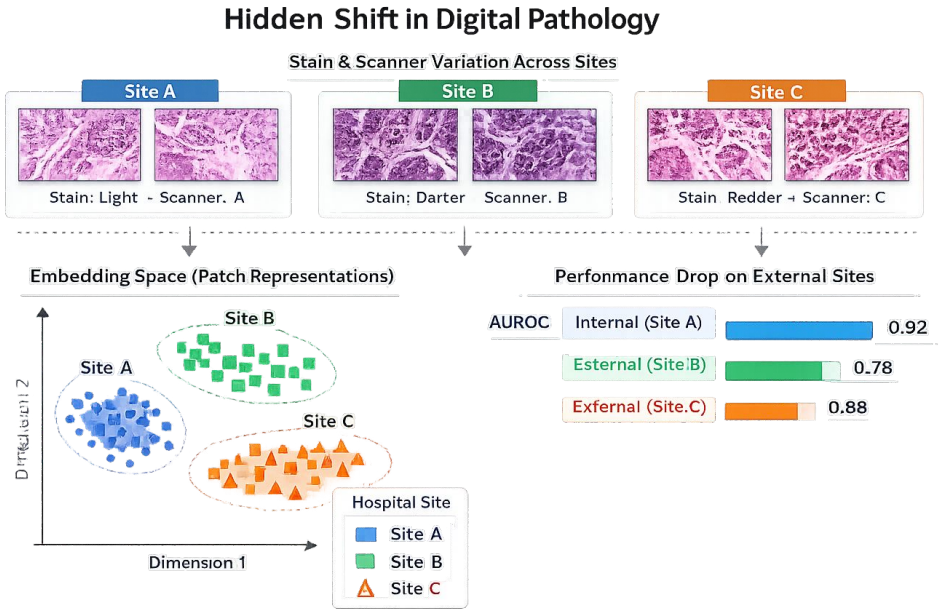


Why Generalization is Difficult in Pathology

Sources of Hidden Shift

- Stain variation
- Scanner variation
- Tissue processing differences
- Site-specific protocols

Shortcut risk: predicting hospital instead of disease



Generated by ChatGPT-5

Three Papers, One Story

Article | Published: 04 September 2024

A pathology foundation model for cancer diagnosis and prognosis prediction

Xiyue Wang, Junhan Zhao, Eliana Marostica, Wei Yuan, Jietian Jin, Jiayu Zhang, Ruijiang Li, Hongping Tang, Kanran Wang, Yu Li, Fang Wang, Yulong Peng, Junyou Zhu, Jing Zhang, Christopher R. Jackson, Jun Zhang, Deborah Dillon, Nancy U. Lin, Lynette Sholl, Thomas Denize, David Meredith, Keith L. Ligon, Sabina Signoretti, Shuji Ogino, ... Kun-Hsing Yu + Show authors

Nature **634**, 970–978 (2024) | Cite this article

70k Accesses | **385** Citations | **388** Altmetric | Metrics

"General-purpose" = task breadth + external validation

Perspective | Published: 12 April 2023

Foundation models for generalist medical artificial intelligence

Michael Moor, Oishi Banerjee, Zahra Shakeri Hossein Abad, Harlan M. Krumholz, Jure Leskovec, Eric J. Topol & Pranav Rajpurkar

Nature **616**, 259–265 (2023) | Cite this article

254k Accesses | **1454** Citations | **751** Altmetric | Metrics

Scale → strong transfer (diagnosis, prognosis)

Article | Published: 19 March 2024

Towards a general-purpose foundation model for computational pathology

Richard J. Chen, Tong Ding, Ming Y. Lu, Drew F. K. Williamson, Guillaume Jaume, Andrew H. Song, Bowen Chen, Andrew Zhang, Daniel Shao, Muhammad Shaban, Mane Williams, Lukas Oldenburg, Luca L. Weishaupt, Judy J. Wang, Anurag Vaidya, Long Phi Le, Georg Gerber, Sharifa Sahai, Walt Williams & Faisal Mahmood

Nature Medicine **30**, 850–862 (2024) | Cite this article

89k Accesses | **990** Citations | **248** Altmetric | Metrics

Generalist medical AI across modalities

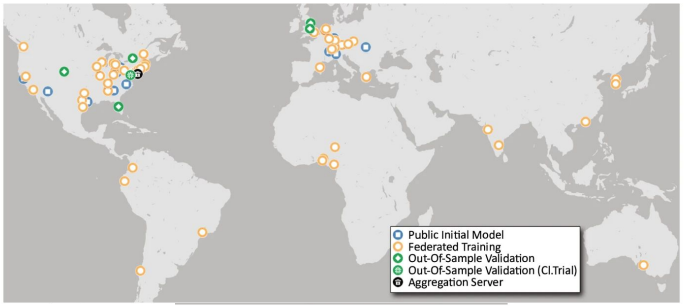
How to Read a “Foundation Model” Paper

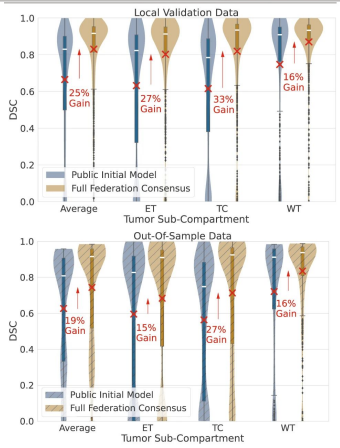
When evaluating claims, check:

- Data scale & diversity
- Pretraining objective
- Transfer protocol
- External validation design

“General-purpose” is an empirical claim

- ☑ Data Scale
- ☑ Objective
- ☑ Transfer
- ☑ External Validation





Pati, S., Baid, U., Edwards, B., Sheller, M., Wang, S. H., Reina, G. A., ... & Poisson, L. (2022). Federated learning enables big data for rare cancer boundary detection. Nature communications, 13(1), 7346.

Lecture Outline

1. ~~Foundation Models~~
2. ~~Pathology Foundational Models~~
3. Model Interpretability
4. Model Generalizability

Why Interpretability Exists

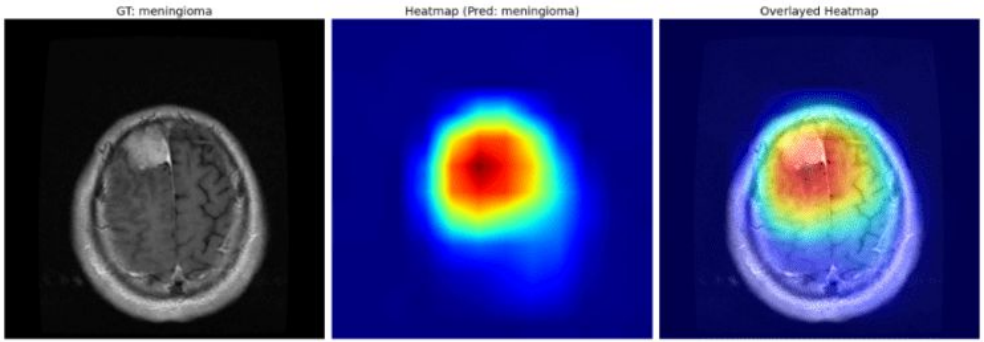
Interpretability
Methods that help humans understand
what a model uses to make predictions

Why it exists:

- Debugging & failure detection
- Communication & trust

Not proof of causality

Keep it clean. No more than that.



https://learnopencv.com/intro-to-gradcam/

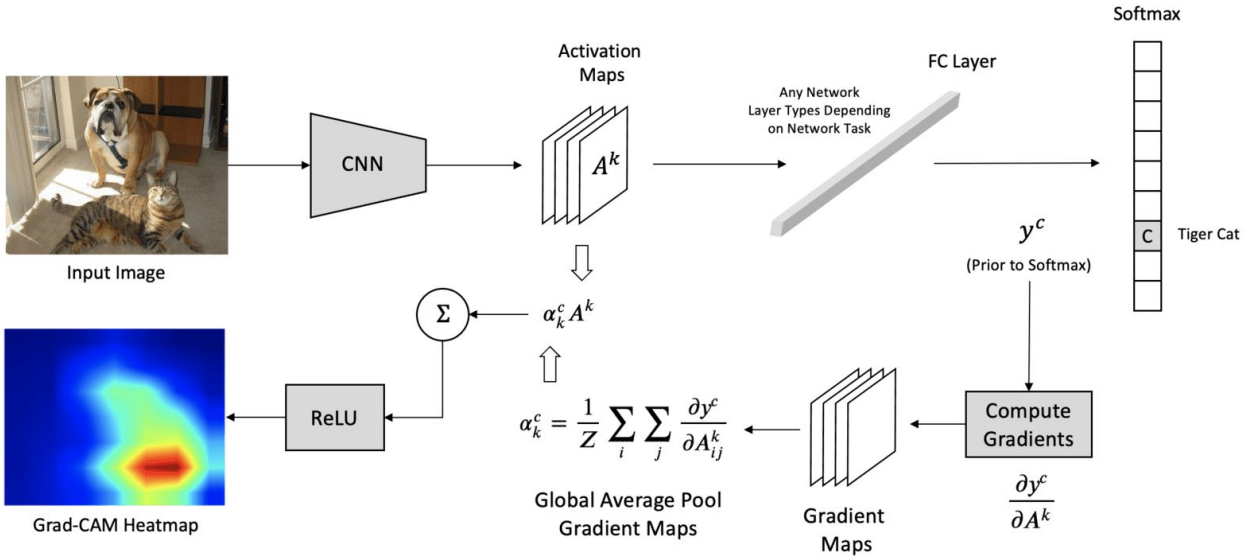
Attribution Methods: What They Show (and What They Don't)

Common attribution tools:

- Saliency / Grad-CAM
- Attention maps
- Occlusion tests

They measure **prediction sensitivity**, not biological reasoning.

Keep that last line visually separated.



Draelos, R. L., & Carin, L. (2020). Use HiResCAM instead of Grad-CAM for faithful explanations of convolutional neural networks. arXiv preprint arXiv:2011.08891.

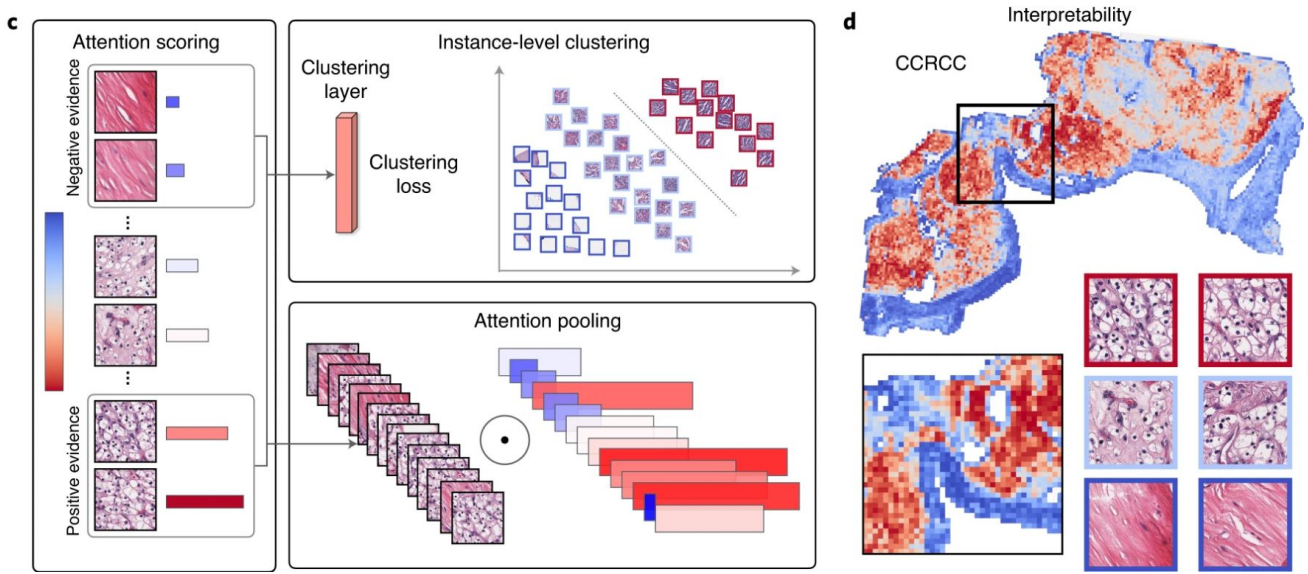
Example — Attention-Based MIL in Pathology (CLAM)

Clustering-constrained Attention MIL (CLAM)

- Attention-based multiple instance learning
- Identifies discriminative regions
- Enables weakly supervised localization

Attention ≠ causation

Keep it tight. Let the figure do most of the work.



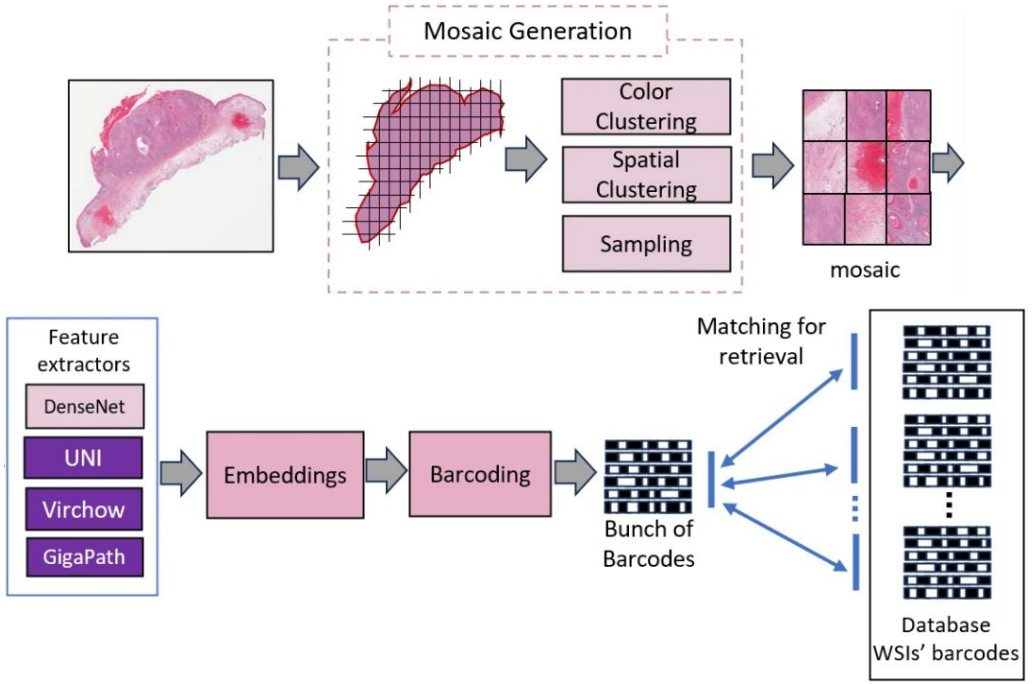
Lu, M. Y., Williamson, D. F., Chen, T. Y., Chen, R. J., Barbieri, M., & Mahmood, F. (2021). Data-efficient and weakly supervised computational pathology on whole-slide images. Nature biomedical engineering, 5(6), 555-570.

Retrieval-Based Interpretability in Pathology

Zero-shot whole-slide retrieval from learned embeddings

Retrieval provides *case evidence*, but:

Similarity ≠ diagnosis



Alfasly, S., Alabtah, G., Hemati, S., Kalari, K. R., Garcia, J. J., & Tizhoosh, H. R. (2025). Validation of histopathology foundation models through whole slide image retrieval. Scientific reports, 15(1), 3990.

Lecture Outline

1. ~~Foundation Models~~
2. ~~Pathology Foundational Models~~
3. ~~Model Interpretability~~
4. Model Generalizability

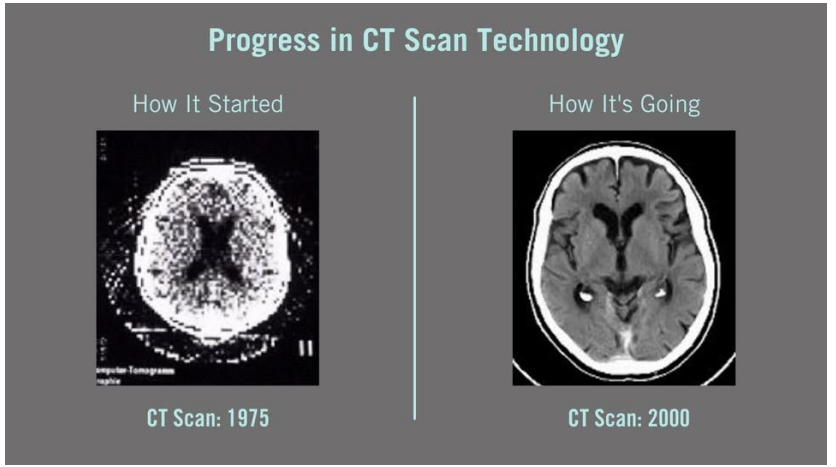
Why Models Fail Outside the Lab

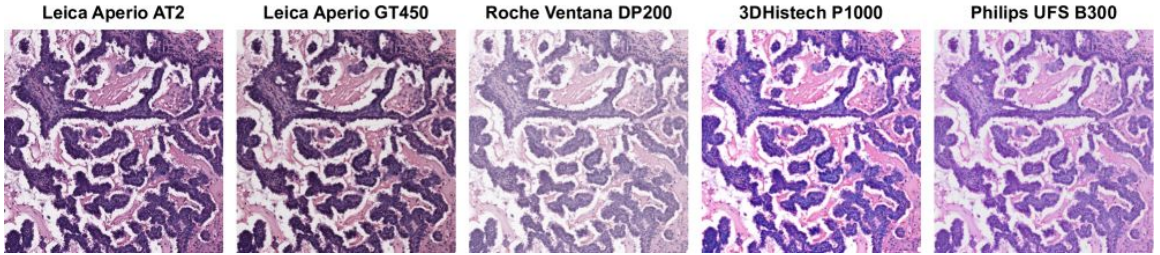
Distribution shift is the norm

Common shifts:

- Site / scanner / protocol
- Temporal
- Population
- Label & prevalence

Shortcut learning amplifies shift





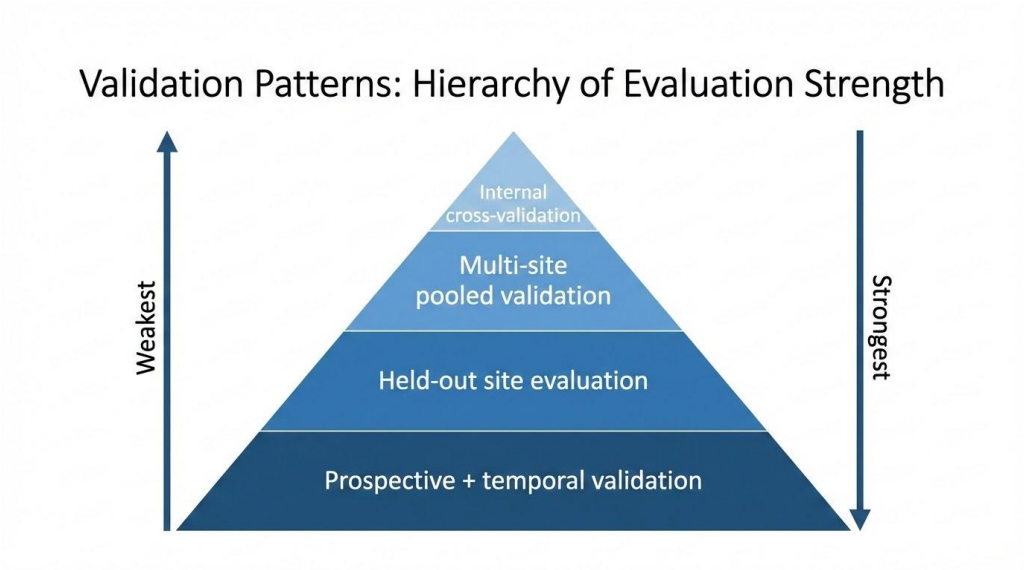
Ryu, J., Song, H., Lee, S., Cho, S. I., Shin, J., Paeng, K., & Pereira, S. (2025, September). SCORPION: Addressing Scanner-Induced Variability in Histopathology. In International Workshop on Uncertainty for Safe Utilization of Machine Learning in Medical Imaging (pp. 102-111). Cham: Springer Nature Switzerland.

What Robust Evaluation Actually Requires

Robust evaluation includes:

- External site holdout
- Temporal split
- Subgroup performance
- Calibration under shift

Stress tests > average AUROC



Federated Learning

Motivation

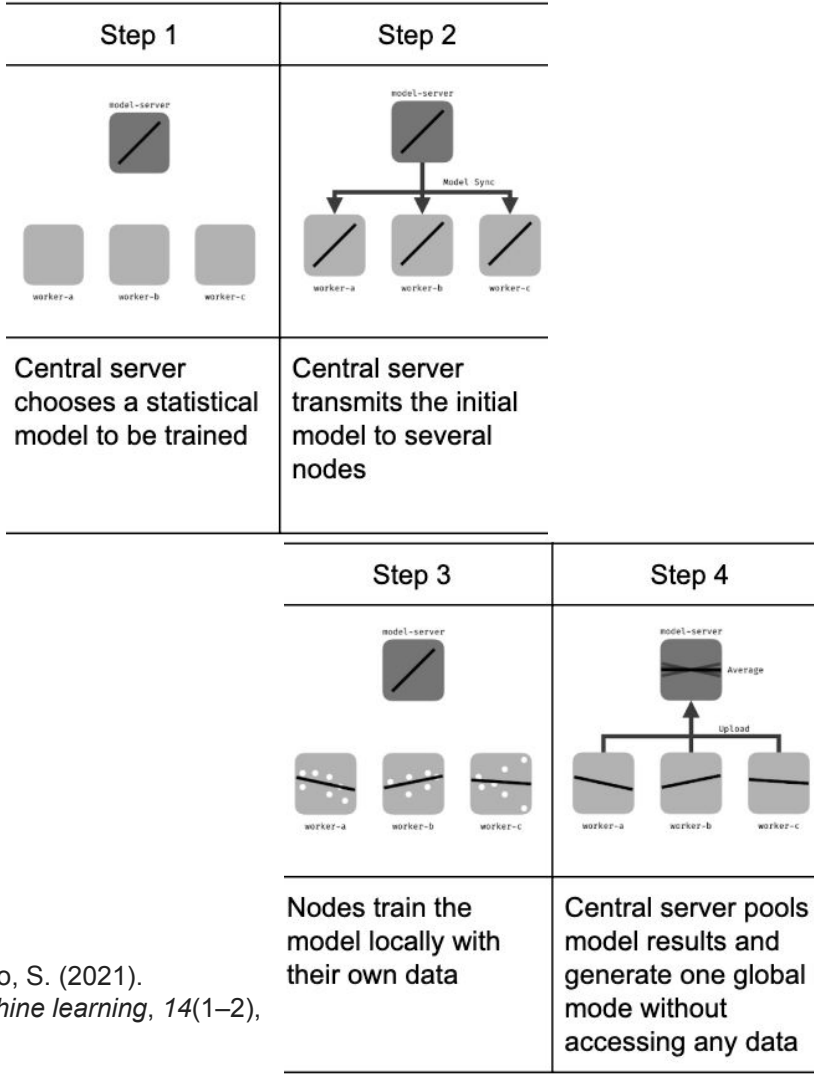
- Train collaboratively across hospitals without sharing raw data
- Larger effective dataset, privacy preserved

Challenges

- Communication overhead, model synchronization
- Data heterogeneity (different scanners, protocols)

Potential Impact

- Improved model generalizability
- Regulatory compliance (HIPAA, GDPR)



Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., ... & Zhao, S. (2021). Advances and open problems in federated learning. *Foundations and trends® in machine learning*, *14*(1–2), 1-210.

Federated Learning

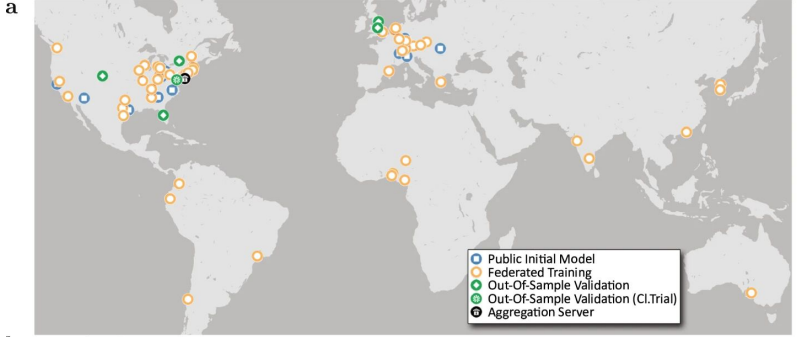
- **Largest FL study in medical imaging → 71 sites, 6 continents**
- **6,314 cases** → Largest glioblastoma dataset
- **No data sharing** → Privacy-preserving model training

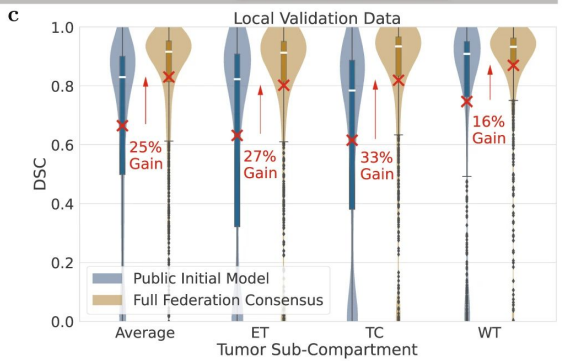
Key Results

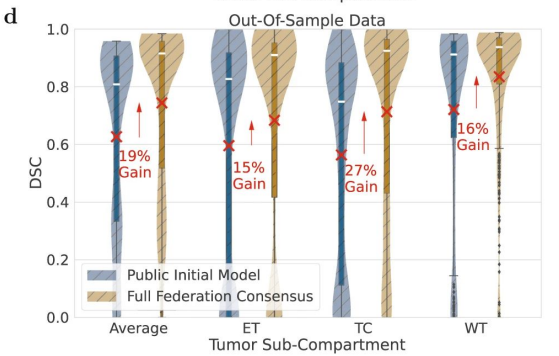
- **+33% improvement** in surgically targetable tumor segmentation
- **+23% improvement** in complete tumor segmentation
- **Validated on:**
  - **Local site data (n = 1,043 cases)**
  - **Out-of-sample data (n = 518 cases)**

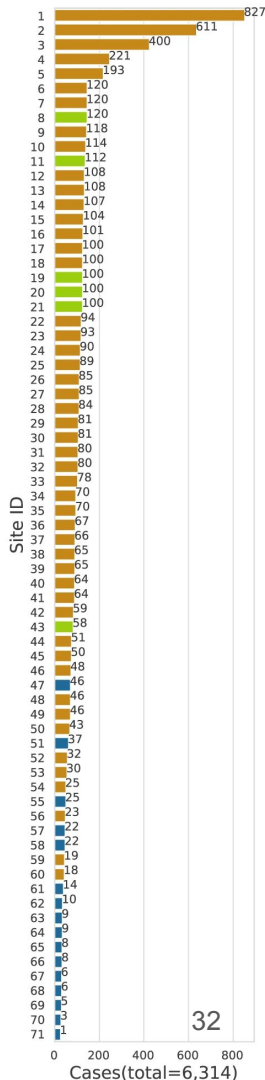
Impact

- **More diverse, generalizable AI models**
- **Public release of the consensus model**
- **New standard for multi-site AI training**









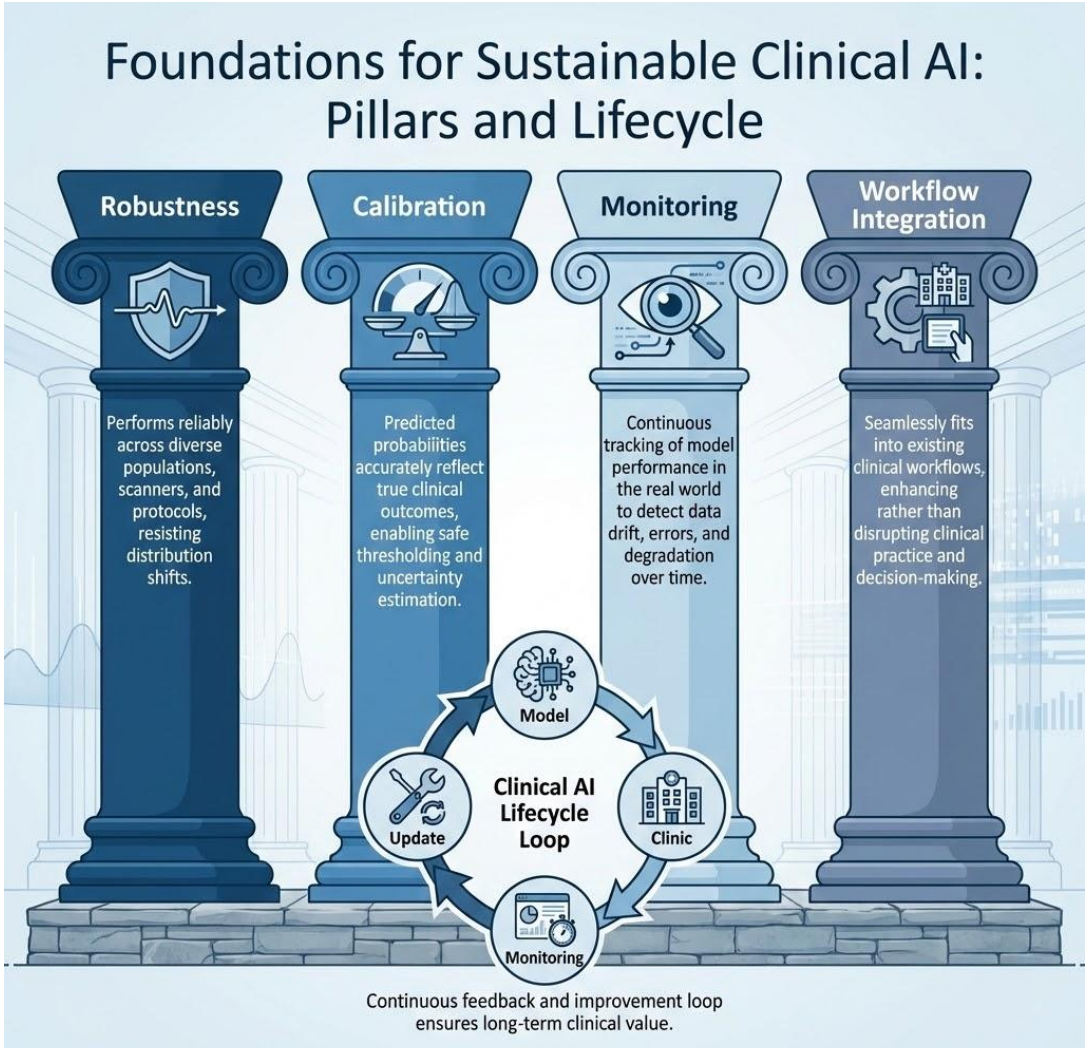
Pati, S., Baid, U., Edwards, B., Sheller, M., Wang, S. H., Reina, G. A., ... & Poisson, L. (2022). Federated learning enables big data for rare cancer boundary detection. *Nature communications*, *13*(1), 7346.

The Real Bar for “Generalist” AI

A generalist model must be:

- Stable across sites
- Calibrated under shift
- Transparent about failures
- Continuously monitored

Representation quality ≠ clinical safety



Recht et al., ICML 2019
Finlayson et al., NEJM 2021
Wynants et al., BMJ 2020

The End