AI-Based Histopathology for Lung Cancer Diagnosis

Omar Graia, Katrina Liu, Audrey Wang

PRIMES Mentor: Dr. Marly Gotti

SWIM 2025

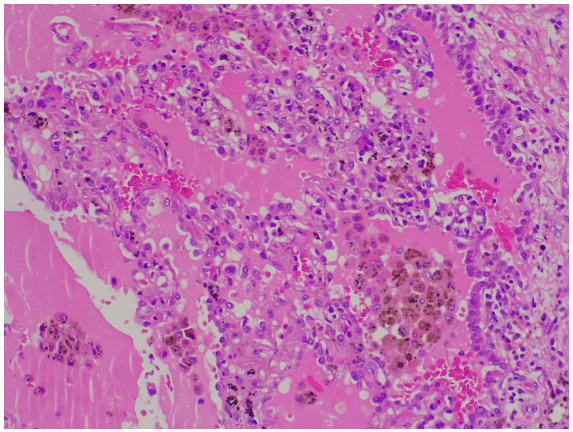
August 21

Outline

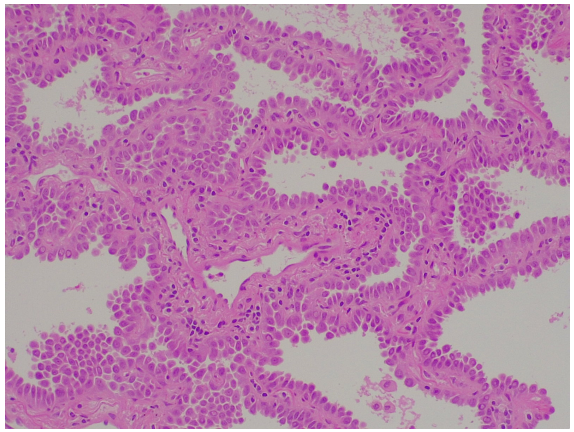
Introduction

- **AI in Healthcare:** Transforming diagnostics and decision-making by analyzing complex medical data.
- **Histopathology:** Microscopic study of tissue samples to diagnose disease, essential for cancer detection.
- **Medical Imaging:** Automated analysis of histopathology images improves accuracy, consistency, and efficiency.
- **Lung Cancer:** Among the most lethal cancers worldwide, requiring rapid and precise diagnosis.
- **Traditional Pathology:** Manual assessments are accurate but time-intensive and subject to variability.
- **Our Focus:** Evaluate AI-based methods on the **LungHist700 dataset**, emphasizing:
  - Robustness across datasets
  - Multi-scale learning
  - Explainable AI

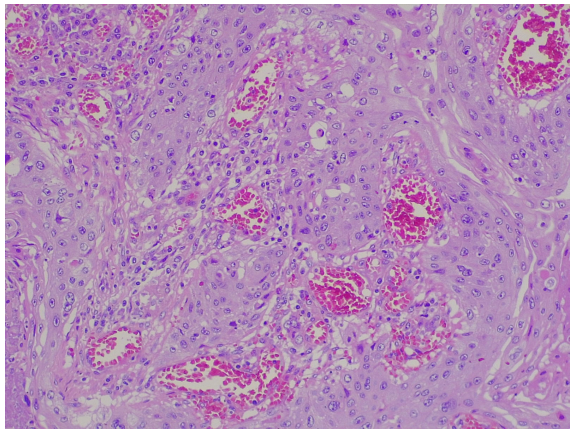
Example Lung Tissue Images



Normal (nor, 20×)



Adenocarcinoma (aca, 20×)



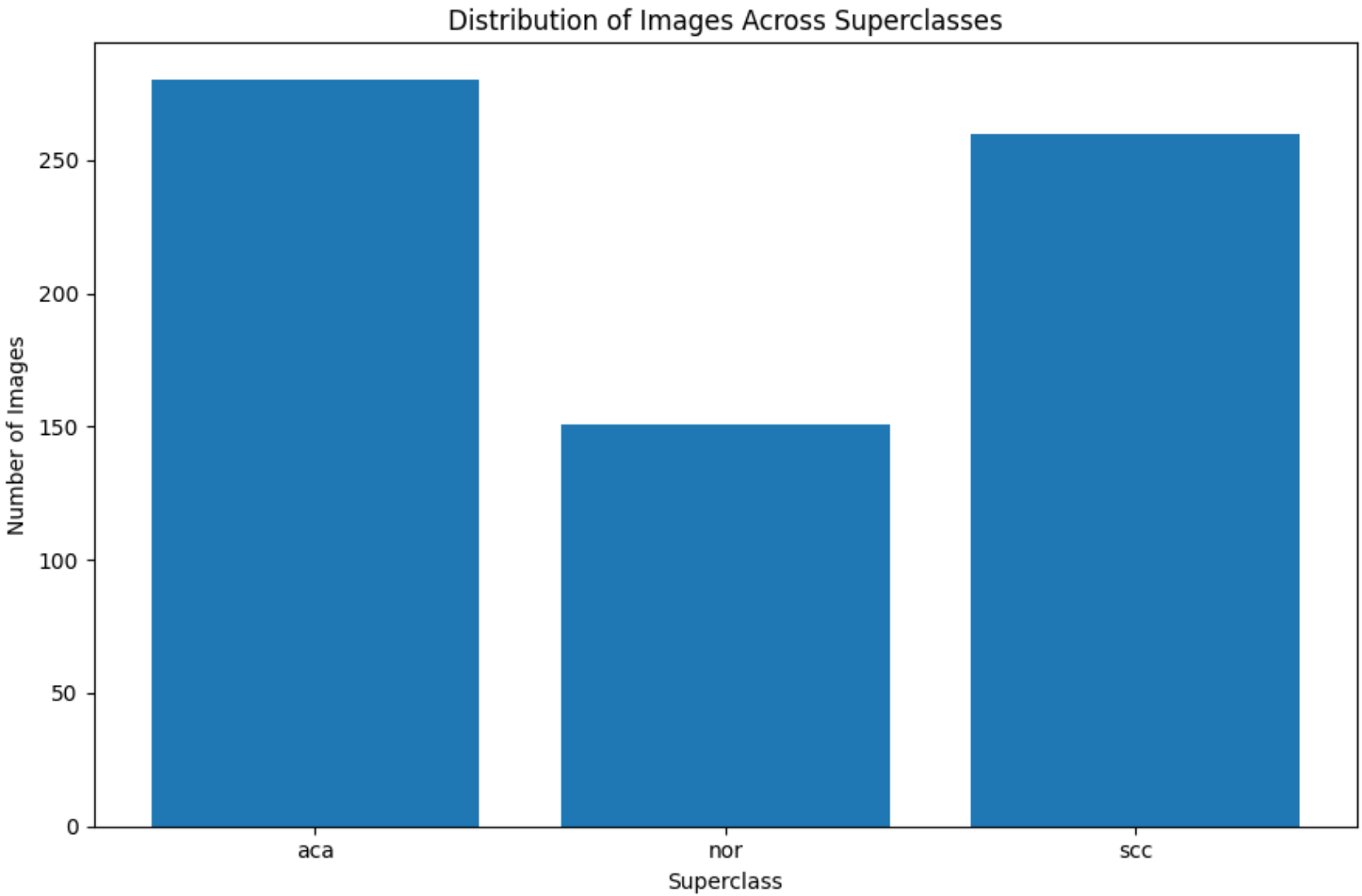
Squamous cell carcinoma (scc, 20×)

Datasets

Dataset	Patients	Labels	Magnif.	Native
LungHist700	**45**	*nor*, *aca*, *scc*	20×, 40×	1200 × 1600 RGB
LC25000	—	*nor*, *aca*, *scc*	(n/a)	patch tiles (RGB)

- LungHist700 has per-patient IDs ⇒ strict patient-wise splits. LC25000 has *no* patient identifiers.
- Many tiles in LC25000 are augmented; naive random splits can include *near-duplicates* ⇒ **data leakage**.
- Train/validate on LungHist700 then evaluate on LC25000 with identical preprocessing.

Distribution of LungHist700 Classes



Key Research Questions (1/2)

1. **Cross-Dataset Robustness**: How well do LungHist700-trained models generalize to external datasets (e.g., LC25000) with different magnifications and staining protocols?
2. **Subtype Classification**: Can models accurately distinguish adenocarcinoma subtypes: poorly, moderately, and well differentiated?

Key Research Questions (2/2)

1. **Multi-Scale Learning**: Does combining 20× and 40× magnifications improve classification performance on LungHist700?
2. **Explainable AI**: What insights do XAI methods (e.g., Grad-CAM, SHAP) provide into model decision-making on lung tissue classification?

What is Logistic Regression?

- **Purpose:** A simple model for binary classification (e.g., normal vs. cancerous tissue).
- **How it works:**
  - Computes a weighted sum of input features.
  - Applies the **sigmoid function** to map values into [0,1].
  - Output interpreted as probability of class $= 1$.
- **Decision rule:** Classify as 1 if $P(y = 1|x) > 0.5$, otherwise 0.

Logistic Regression

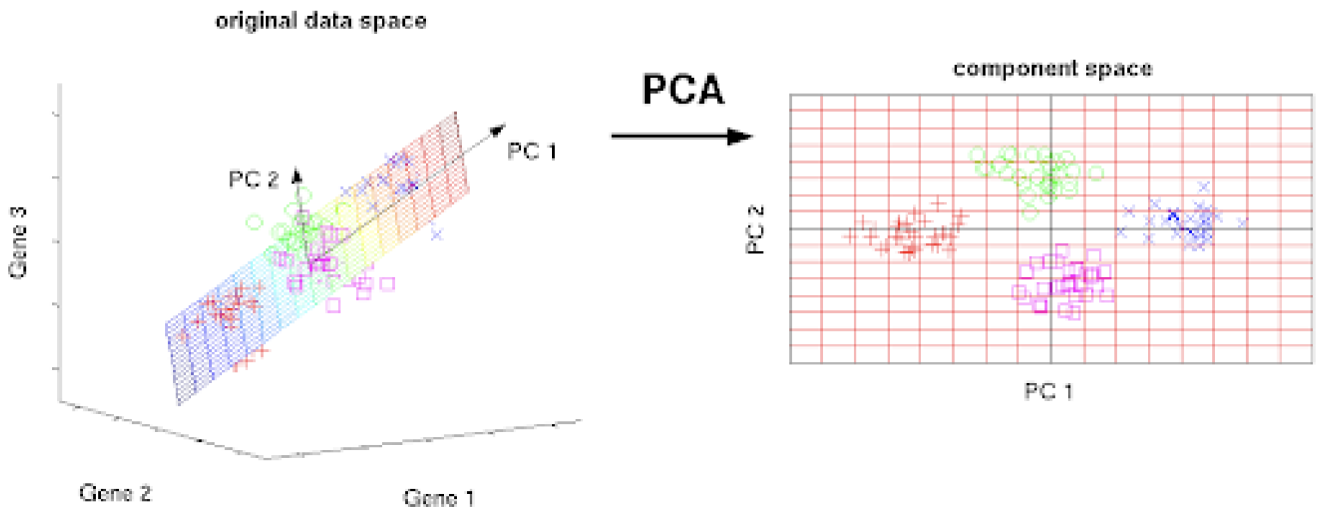
- **Pros:** Interpretable, efficient, widely used.
- **Cons:** No spatial bias, may overfit in high-dimensional spaces.

Logistic Regression for Binary Classification

- Task: Distinguish between **normal** vs. **cancerous** lung tissue.
- Dataset: LungHist700
- Preprocessing:
  - Downsampled images: $1200 \times 1600 \rightarrow 30 \times 40$ pixels
  - Flattened into 3600-D RGB vectors
  - Labels: **normal**, **cancerous** (including *adenocarcinoma (aca)*, *squamous cell carcinoma (scc)*)
- Dimensionality reduction with PCA (95% variance retained).

Principal Component Analysis (PCA)

- **PCA** is a dimensionality reduction technique.
- Preserves as much variance as possible.
- It finds new orthogonal axes (principal components) that capture the most variance in the data.
- Keeps only the top components that explain most of the variance, reducing noise and redundancy.
- Make training more efficient and stable.



Patient-level Sampling Strategy

- **20 patients total**
  - Balanced design: 10 normal, 10 cancer (5 aca, 5 scc)
  - Reflects real-world constraint of limited tissue samples
- **One image per patient**
  - Several Normal patients have only one image
  - Prevents patients with many images from dominating training
  - Ensures equal contribution across patients
  - Avoids patient-specific feature leakage

Evaluation Protocol

1. Test set 1: All remaining images (including extra images from training patients). Total 671 images.
2. Test set 2: All images from 25 unseen patients. Total around 400 images.

Logistic Regression Models

Model 1

Raw pixel features (no PCA)

Model 2

PCA-transformed features (95% variance)

Model 3

PCA + full LungHist700 (20× magnification) → LC25000

Model 4

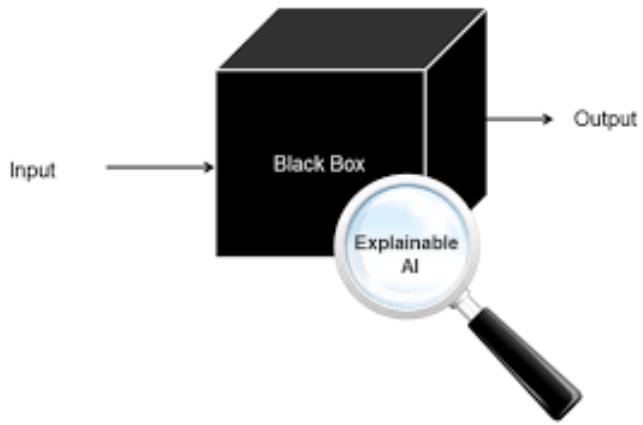
PCA + full LungHist700 (40× magnification) → LC25000

Cross-Dataset Generalization

- Models 3 & 4 were trained on the entire LungHist700 dataset.
- Evaluated on 600 LC25000 images:
  - 200 adenocarcinoma
  - 200 squamous cell carcinoma
  - 200 normal
- Images resized and preprocessed consistently.

Explainable AI: Why It Matters

- Many ML models act as **black boxes**.
- Explainable AI (XAI) provides transparency in decision-making.
- Helps answer:
  - Why did the model make this prediction?
  - Which features (pixels) influenced the decision most?
  - Are predictions grounded in medically relevant regions?
  - Model Debugging



Explainable AI: Clinical and Regulatory Importance

- Critical in medicine: false positives = unnecessary interventions, false negatives = missed diagnoses.
- Allows clinicians to validate predictions and override when needed.
- In practice: transparency is essential for acceptance in real diagnostic workflows.

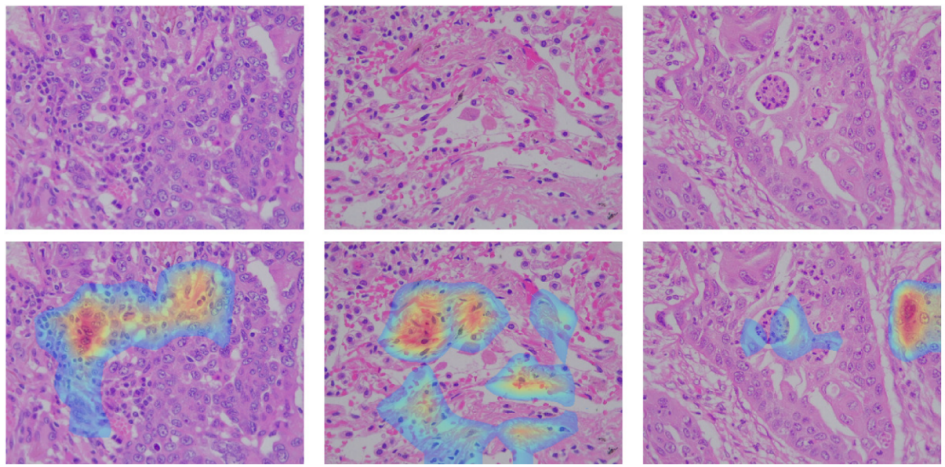


Figure: XAI Example

- **Tool we used:** SHAP (SHapley Additive exPlanations).

What is SHAP?

- **SHAP = SHapley Additive exPlanations**
- A method to explain predictions of machine learning models.
- Based on **Shapley values** from cooperative game theory:
  - Treats each feature as a "player" in a game.
  - Calculates how much each feature contributes to the prediction.
- Consistent: features that matter more get higher importance.

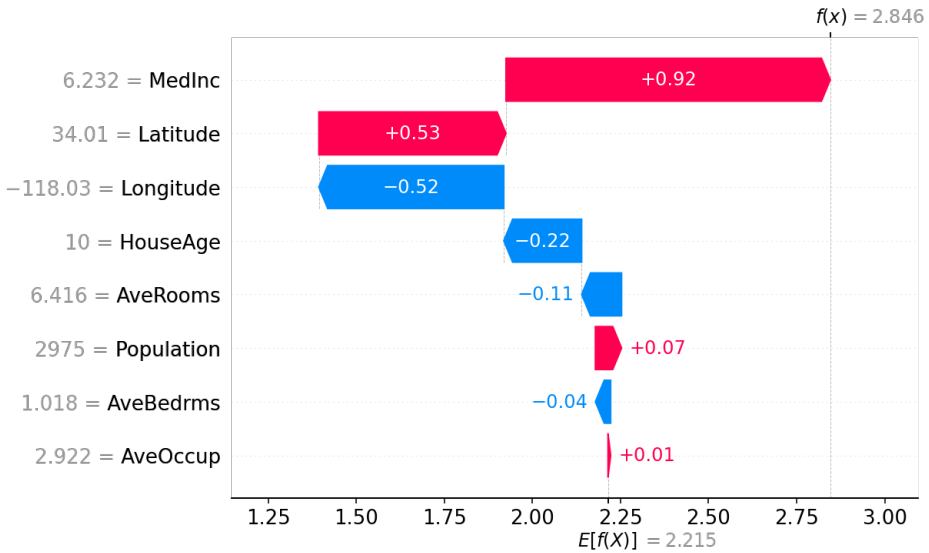


Figure: Housing Example

SHAP in Our Study

- Applied SHAP to **Model 1** (raw pixel logistic regression).
- PCA was **excluded** for interpretability:
  - PCA produces abstract components.
  - Cannot directly map attributions back to pixels.
- SHAP heatmaps computed on resized 30×40 pixel images.
- Revealed spatial patterns used by the model for classification.
- Enabled checking if predictions relied on medically meaningful regions.

What is a CNN?

- CNN stands for Convolutional Neural Network.
- Special type of deep neural network for image recognition and classification.
- Learns hierarchical patterns through convolutional layers.
- Effective in tasks involving spatial or visual data.

How Does a CNN Work?

- **Convolution Layers**: Detect local spatial features.
- **Pooling Layers**: Downsample to reduce spatial resolution.
- **Fully Connected Layers**: Combine features for classification.
- **Activation Functions**: Introduce non-linearity (e.g., ReLU).

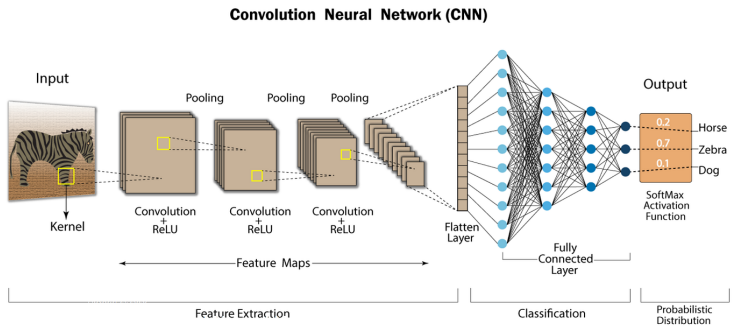


Figure: Typical CNN structure: conv + pool + FC layers

Different Types of CNN Architectures

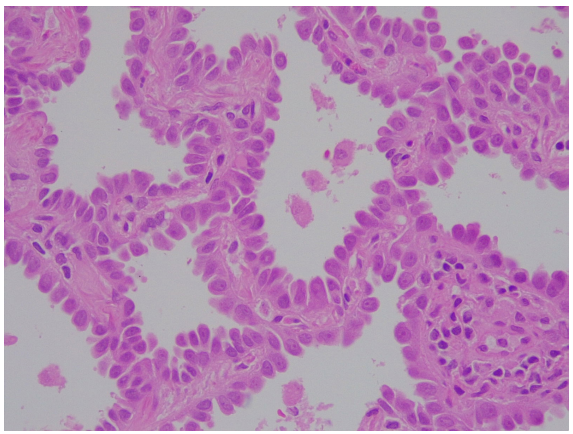
- **LeNet-5**: Early CNN for digit recognition.
- **AlexNet**: Pioneered deep CNNs.
- **VGGNet**: Uses many small (3x3) filters.
- **ResNet**: Introduces skip connections for deeper networks.

ResNet50V2

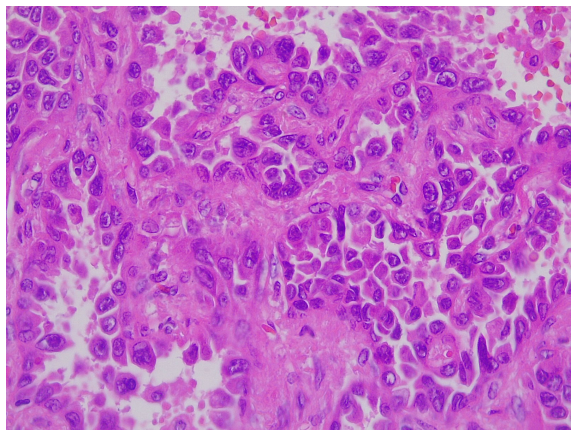
- 50-layer deep residual network.
- Uses identity shortcut connections to address vanishing gradients.
- Variant V2 improves pre-activation and training stability.
- Widely used for medical and natural image classification.

Adenocarcinoma Differentiation

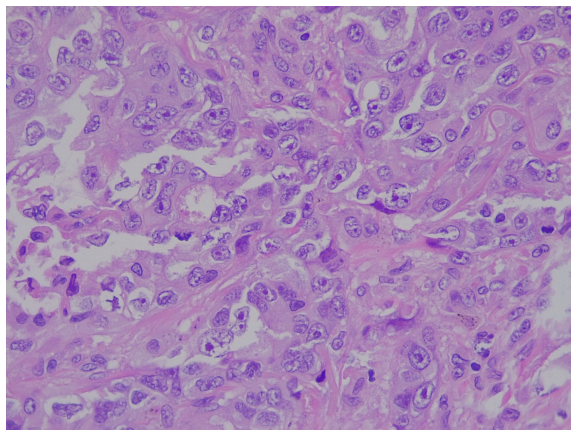
- Task: classify adenocarcinoma (aca) into:
  - Well-differentiated
  - Moderately differentiated
  - Poorly differentiated
- Model: ResNet-50 backbone with custom 3-class head.
- Data: LungHist700 (40× images).
- Evaluation: three patient-level train/test splits to reduce leakage.



Well-differentiated



Moderately-differentiated



Poorly-differentiated

Methods: LC25000 Generalization

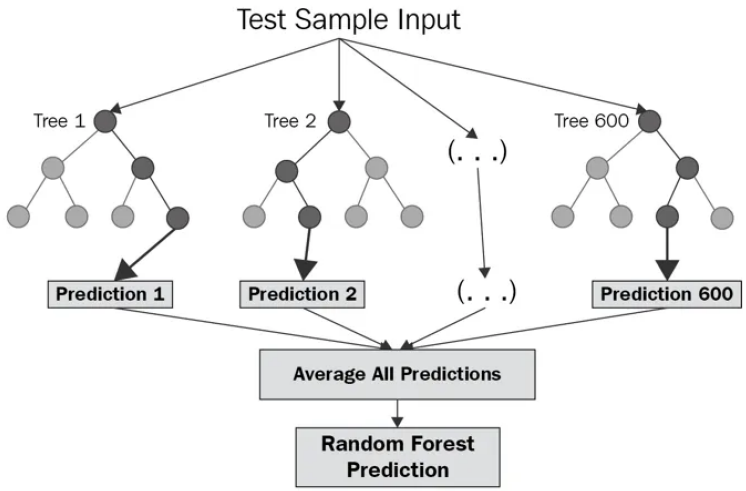
- Models trained on LungHist700, evaluated on external LC25000 dataset.
- Models considered:
  - Logistic Regression (HOG features, 40×)
  - Random Forest (20× vs. 40× with PCA)
  - ResNet50V2 (40× images, ImageNet pre-trained)
- Evaluation: 300 LC25000 images (balanced across *nor*, *aca*, *scc*).
- Identical preprocessing pipelines across models to ensure comparability.

Overview of Multi-Scale Learning Approach

- **Goal:** Assess whether combining magnifications (20× and 40×) improves classification of *nor*, *aca*, *scc* on LungHist700.
- **Baseline vs. Multi-Scale:** Baseline uses only 20×; Multi-scale trains on both 20× and 40×.
- **Rationale:** Complementary spatial detail across scales may capture morphology better than a single scale.
- **External Check:** After training on LungHist700, test generalization on LC25000.

Random Forests

- An ensemble of many decision trees trained on different bootstrap samples (bagging).
- Each split considers a random subset of features, which decorrelates trees.
- Predictions are aggregated: majority vote for classification, mean for regression.
- Strengths: handles nonlinearity, little feature engineering, robust to noise; can provide feature importance.
- Limitations: less transparent than a single tree; larger models and slower inference when many trees.



Baseline 20× & Multi-Scale Model Preprocessing

- Patient-wise split across 45 patients:
  - If a patient has any normal tile ⇒ label normal. If both aca/scc exist, assign the majority class.
  - ∼80/20 train/test split by patient.
- Resize to 100 × 100, RGB, flatten (30,000-dim), standard scaling.
- PCA to reduce dimensionality.
- Random Forest on PCA features.

Improved Feature Extraction for Multi-Scale

- **Color features** (RGB/HSV/CIELAB): mean, std, skewness, kurtosis; per-channel histograms; H&E proxy via blue/red channel stats.
- **Texture features:** GLCM Haralick descriptors at multiple offsets/angles; LBP histograms; Sobel gradients.
- **Morphology:** Otsu thresholding; connected components; top-5 region areas; region counts.
- **Nuclei cues:** Inverted grayscale + Otsu; size-filtered regions; counts and size statistics.
- **Total:** 164 features/image; same patient-wise split and RF classifier.
- **Generalization check:** Evaluate on LC25000.

Texture Features: GLCM & LBP

GLCM (Gray-Level Co-occurrence Matrix):

- Counts how often gray levels i, j occur together at offset (θ, d); normalized matrix $p_{ij}^{(\theta,d)}$.
- Summaries we use: Contrast $= \sum_{i,j}(i-j)^2 p_{ij}$, Energy $= \sum_{i,j} p_{ij}^2$.
- Computed over directions $0°, 45°, 90°, 135°$ and multiple d; concatenated into one feature vector.

LBP (Local Binary Patterns):

- Threshold neighbors vs. center $\Rightarrow$ binary code; histogram over the tile.
- Rotation-invariant, uniform LBP with r=3, P=24; robust to monotone staining shifts.

Otsu Thresholding

- Automatic foreground/background split on a grayscale image.
- Builds a histogram and searches a threshold τ that best separates the two groups (no tuning).
- Produces a binary mask $\Rightarrow$ connected components, region counts, top-k areas, size statistics.
- For nuclei cues: invert grayscale before thresholding; size-filter small artifacts.

Generalized Feature Extraction

- Reduce dataset-specific bias (e.g., staining ratios) to improve cross-dataset generalization.
- Changes:
  - Channel-wise normalization and percentile-based statistics.
  - Multi-scale LBP; gradient-based ratios.
  - Threshold-derived region stats; simplified connectivity metrics.
  - *Removed* dataset-specific H&E ratio terms.
- Model: Same RF setting; comparable feature dimensionality.

Evaluation Strategy

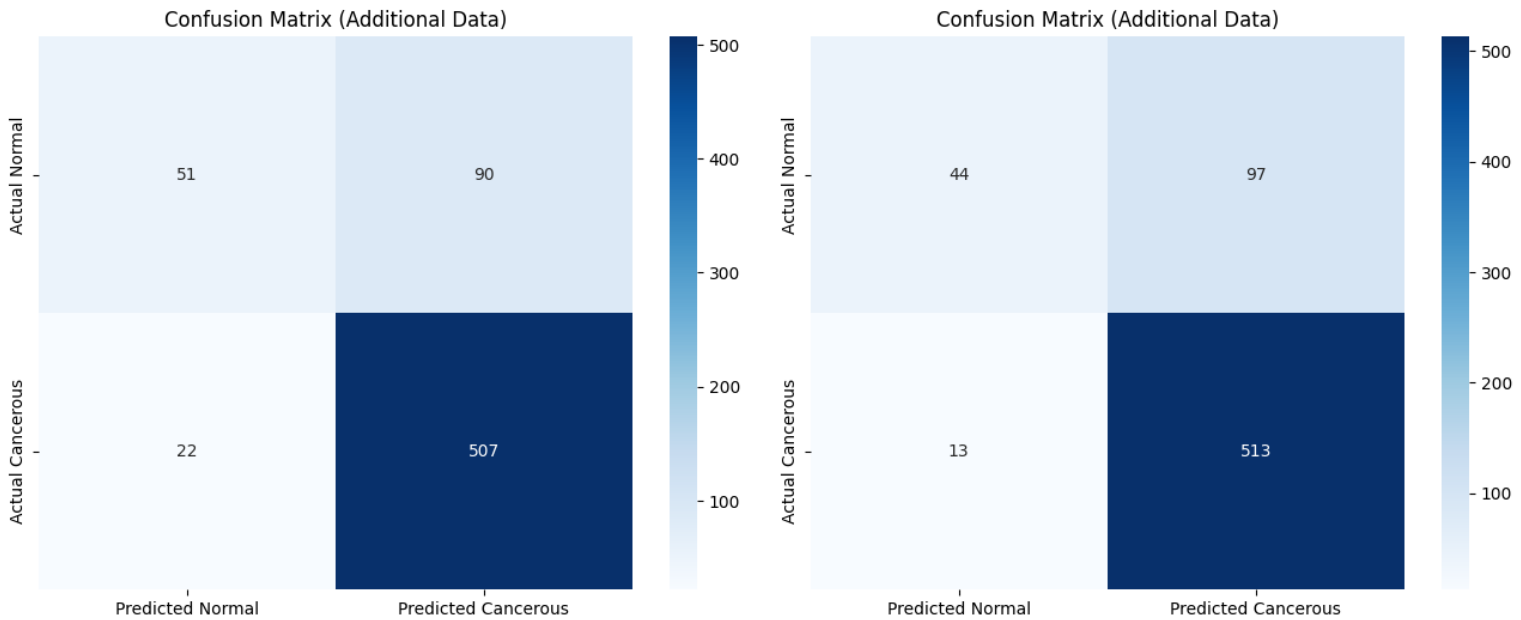
- Two test strategies used:
  1. Remaining 671 images not used for training.
  2. All images from 25 completely unseen patients.
- Models compared:
  - Model 1: Logistic Regression, raw pixel features.
  - Model 2: Logistic Regression, PCA features (95% variance).
  - Model 3: Trained on full LungHist700 (20×), tested on LC25000.
  - Model 4: Trained on full LungHist700 (40×), tested on LC25000.

LungHist700: Models 1 & 2 (All Remaining Images)

Overall better performance with PCA.

Table: Models 1 and 2 on all images of all unseen patients in LungHist700

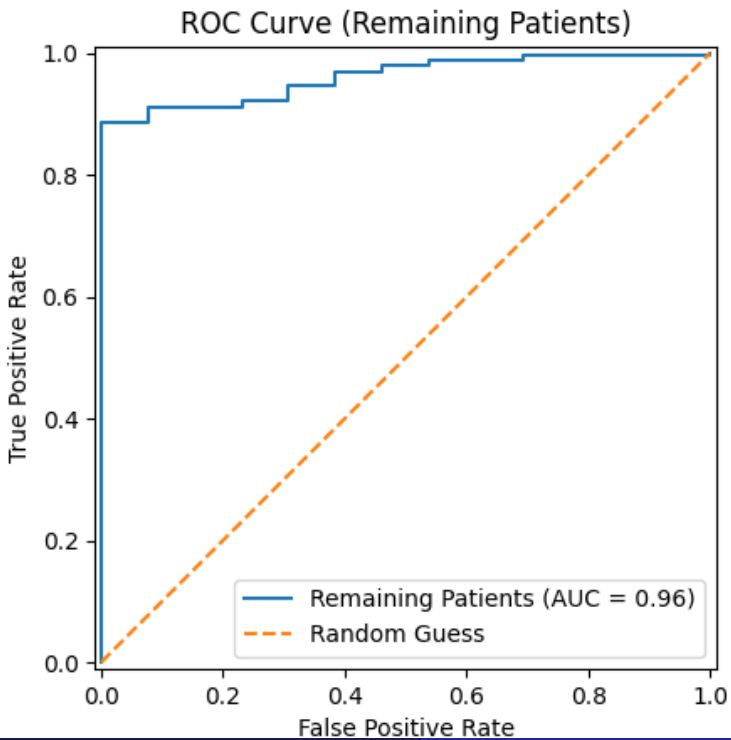
Model	Accuracy	Precision	Recall	f1-score	AUC
Model 1 (No PCA)	0.94	0.94	0.97	0.98	0.96
Model 2 (PCA)	0.96	0.98	0.99	0.98	0.96

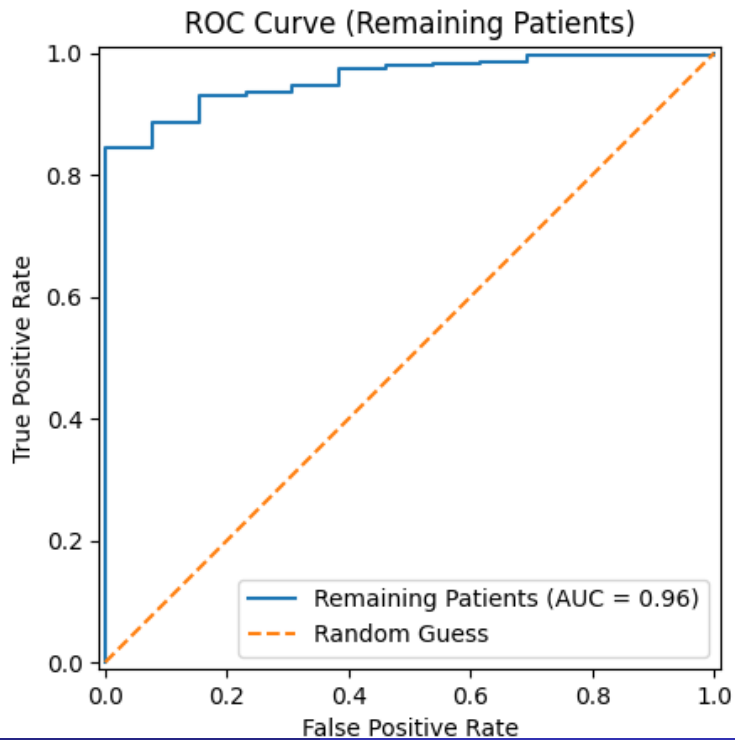


LungHist700: Models 1 & 2 (Unseen Patients)

Table: Performance metrics of models 1 and 2 on all images not used for training in LungHist700

Model	Accuracy	Precision	Recall	f1-Score	AUC
Model 1 (No PCA)	0.82	0.80	0.89	0.94	0.85
Model 2 (PCA)	0.84	0.84	0.98	0.90	0.81

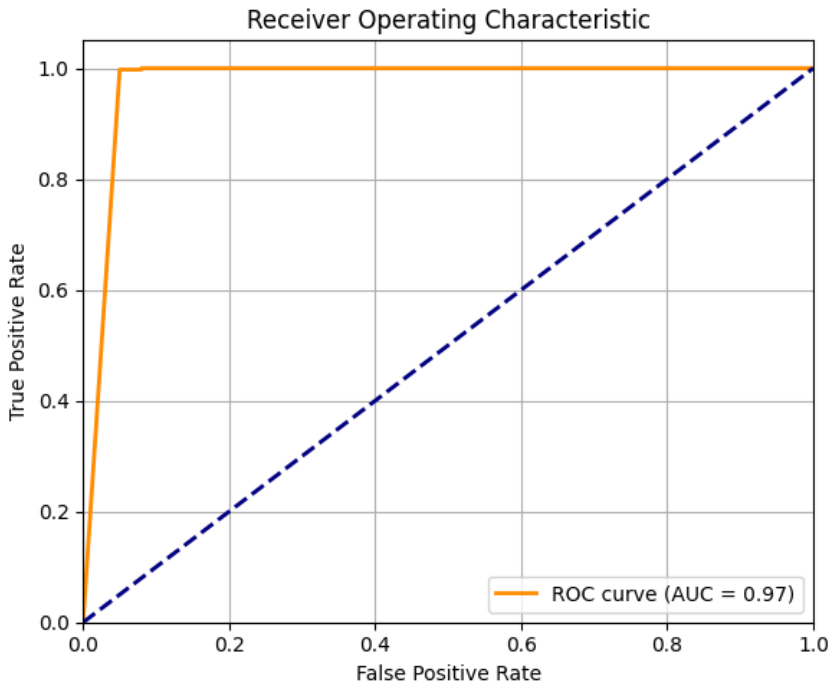


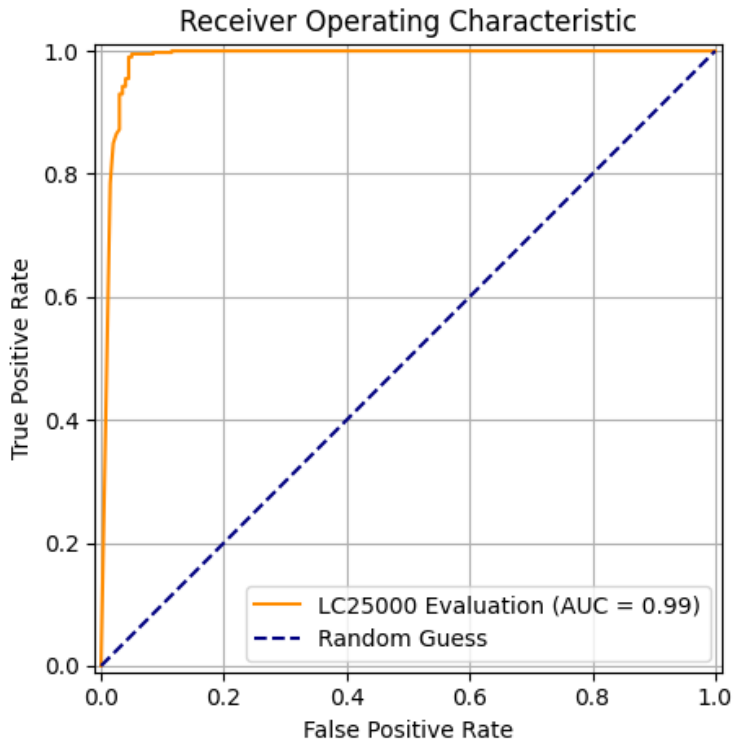


Cross-Dataset Generalization (Models 3 & 4)

Table: Performance metrics of models 3 and 4 on 600 sampled images of LC25000

Model	Accuracy	Precision	Recall	F1-Score	AUC
Model 3 (20 ×)	0.71	0.69	1.00	0.82	0.97
Model 4 (40 ×)	0.80	0.77	1.00	0.87	0.99

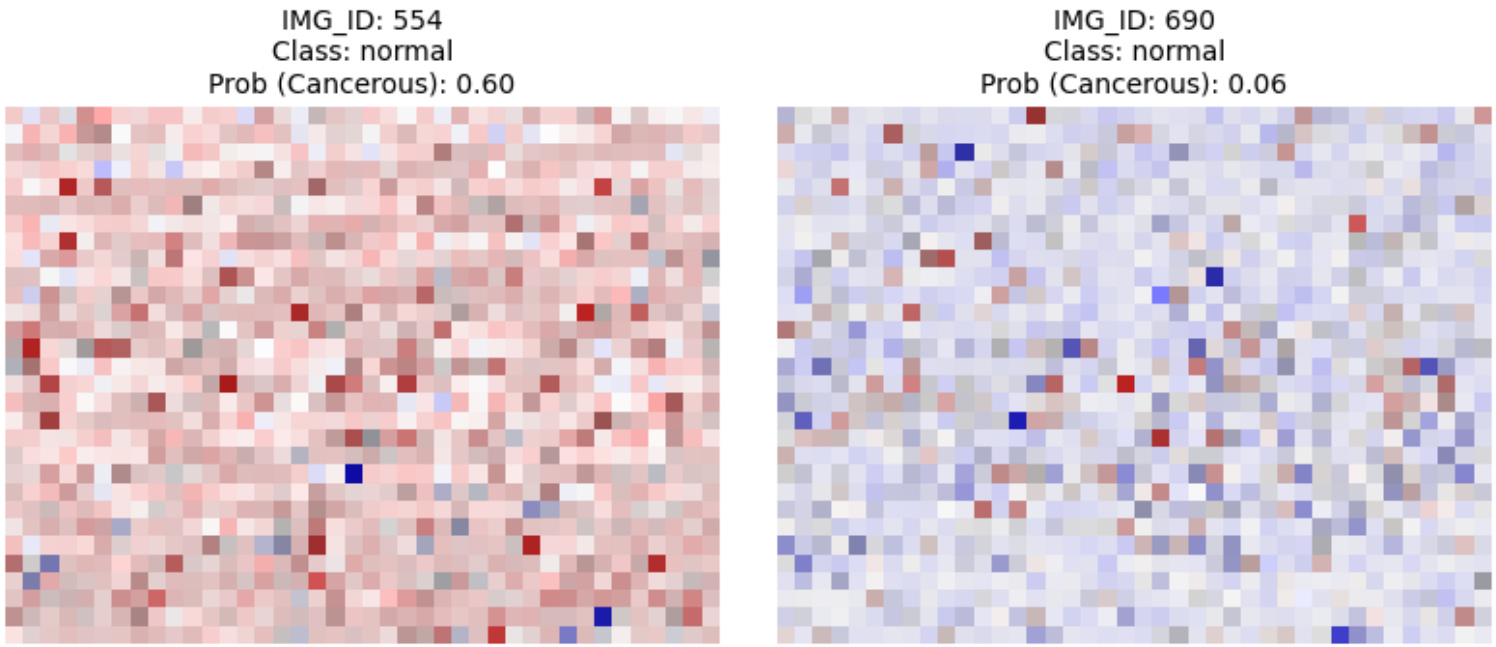




Explainable AI: SHAP Interpretability

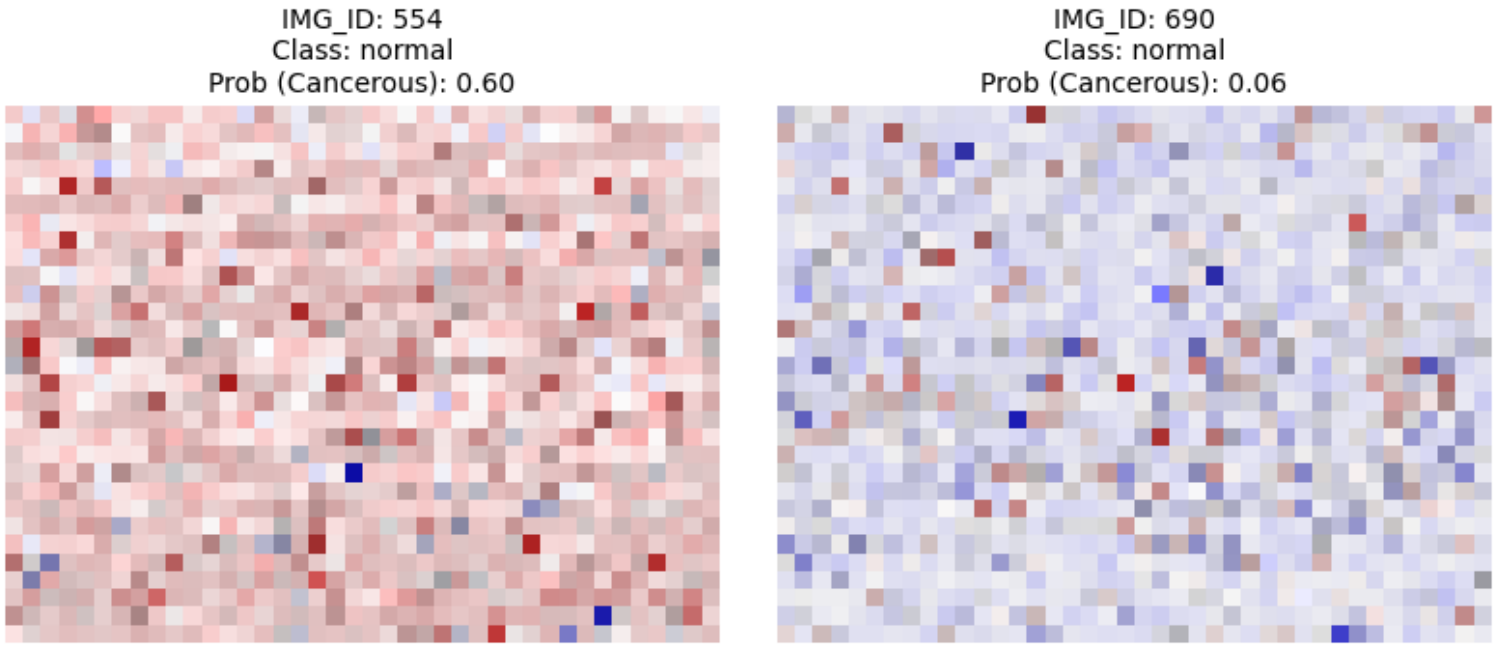
- We use SHAP (SHapley Additive exPlanations) to compute feature-level contributions in lung tissue classification.
- SHAP assigns contribution scores to each pixel.
- Additional processing with heatmap overlay: visualize pixel-wise impact on model predictions.
- BWR colormap:
  - Blue = features pushing prediction toward **normal**
  - Red = features pushing prediction toward **cancerous**
- Enables intuitive interpretation of how models make decisions.
- Applied to Model 1 (raw pixels, trained on 20 LungHist700 images).

SHAP Analysis



- Image ID 690: Mostly blue, predicts **normal** (probability: 0.06).
- Model focuses on outer blue regions, suggesting healthy tissue.
- Central region shows mixed signals, reflecting model uncertainty.

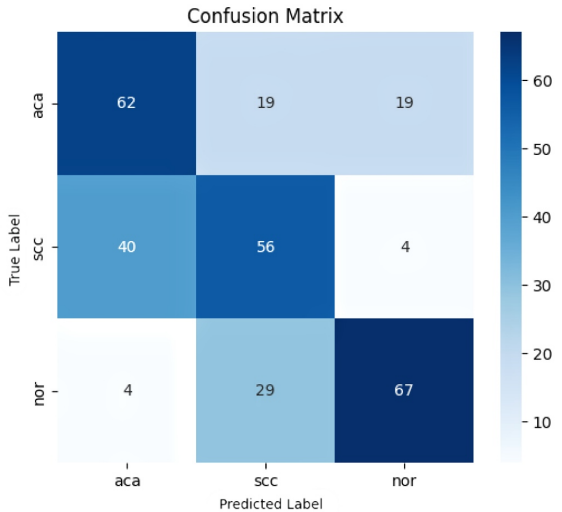
SHAP Analysis



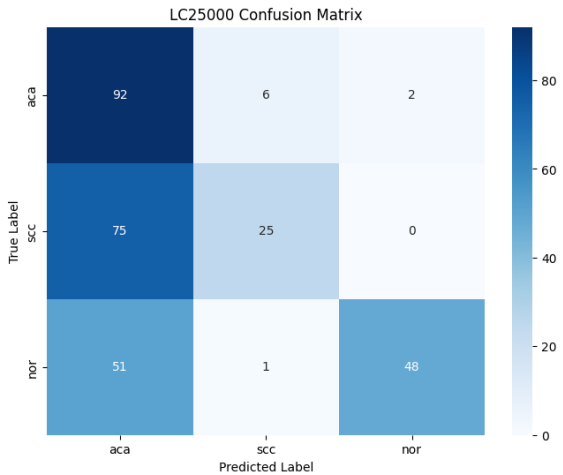
- Misclassified Image ID 554: True label = **normal**, but predicted as cancerous (0.6).
- SHAP reveals red regions due to isolated pixel features, not spatial patterns.
- Highlights logistic regression's limitation: lacks spatial awareness.

LC25000 Generalization: Logistic Regression, Random Forest, ResNet

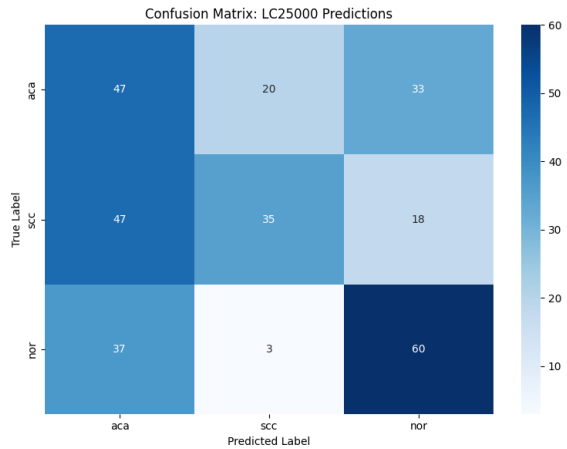
Model	Accuracy	Precision	Recall	F1
ResNet50V2 (40×)	0.62	0.62	0.62	0.62
Random Forest (20×)	0.38	0.57	0.38	0.32
Random Forest (40×)	0.55	0.74	0.55	0.53
Logistic Regression (40×, HOG)	0.47	0.50	0.47	0.47



ResNet50V2 (40×)



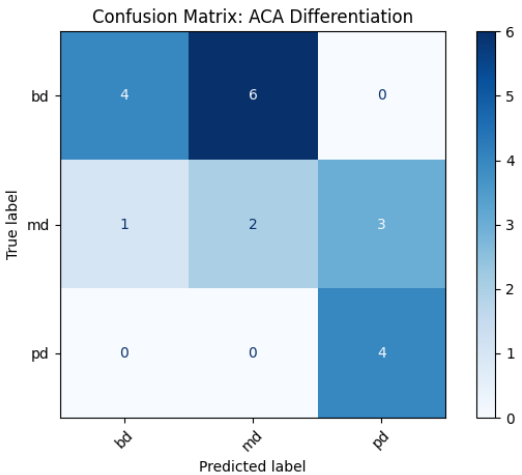
Random Forest (40×)

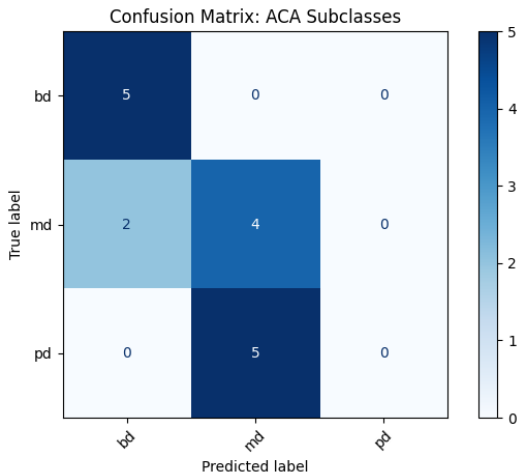


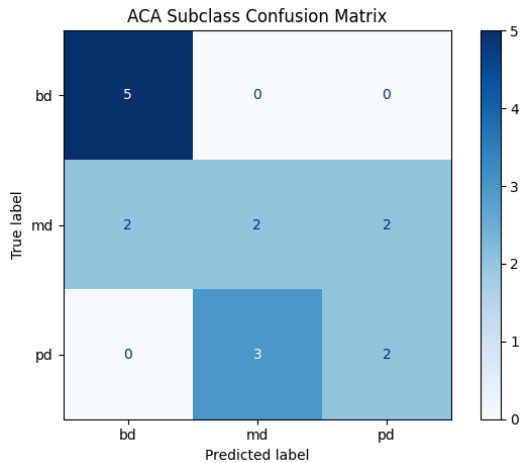
Log Reg (40×)

Differentiation (3 Splits)

- Accuracy range: 0.50 – 0.56
- Split 2 highest = 0.56
- Confusion matrices show overlap between well vs. moderately differentiated.







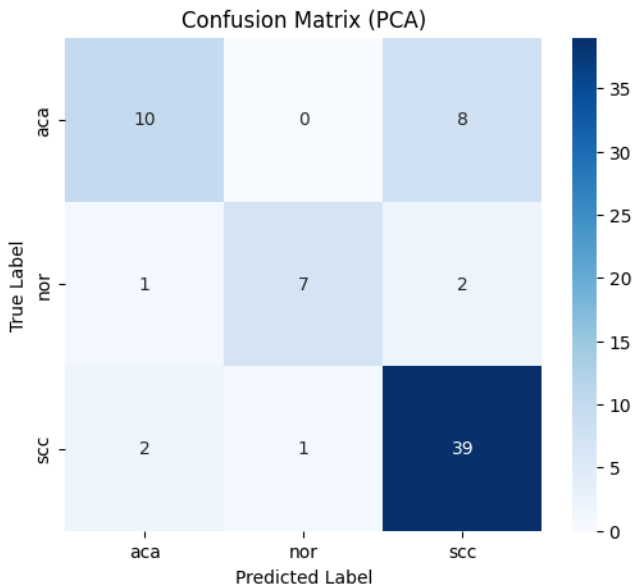
Overall Mutli-Scale Results

	LungHist700				LC25000			
Model	Acc.	Prec.	Rec.	F1	Acc.	Prec.	Rec.	F1
Baseline 20×	0.80	0.80	0.80	0.79	–	–	–	–
Multi-Scale	0.84	0.84	0.84	0.82	0.66	0.75	0.66	0.67
Improved Features	0.95	0.96	0.95	0.95	0.62	0.72	0.62	0.54
Generalized Features	0.86	0.86	0.86	0.86	0.80	0.80	0.80	0.80

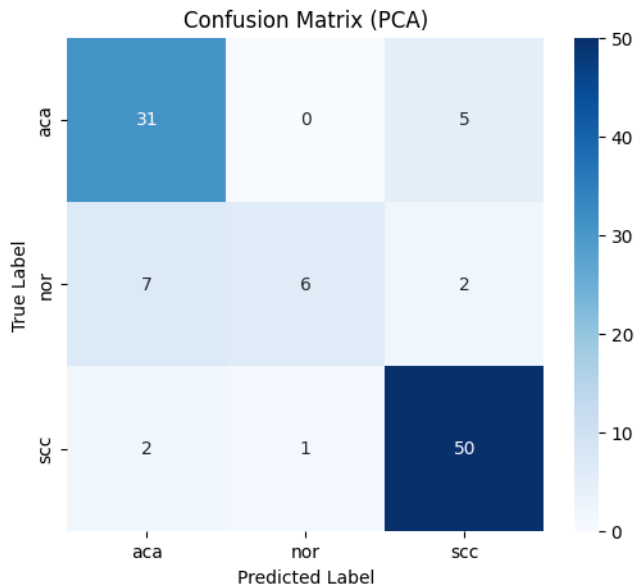
Baseline 20× vs. Multi-Scale

Model	Accuracy	Precision	Recall	F1-Score
Baseline 20×	0.80	0.80	0.80	0.79
Multi-Scale	0.84	0.84	0.84	0.82

Performance on the LungHist700 test set.



Baseline 20×

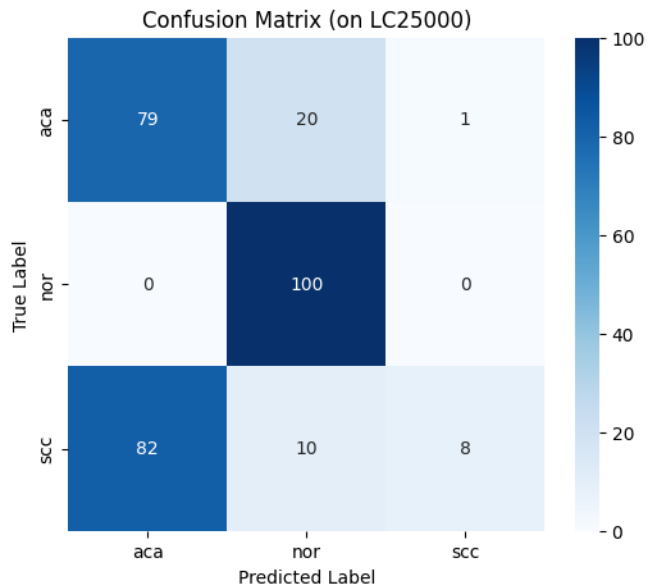


Multi-Scale

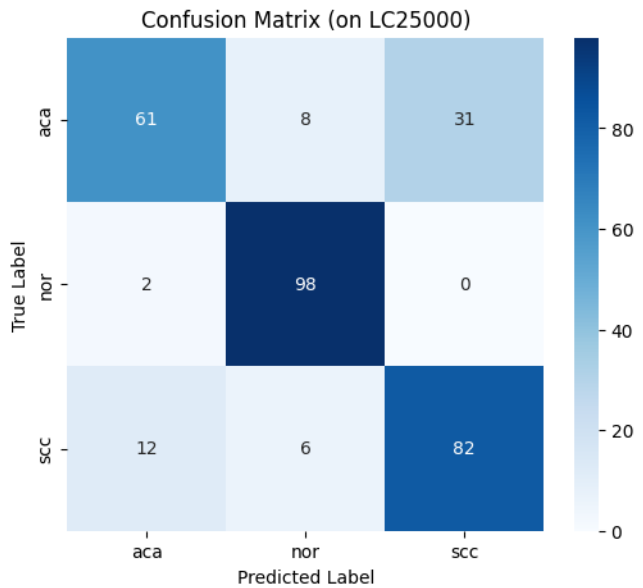
Mutli-Scale Models' Generalization to LC25000

Model	Accuracy	Precision	Recall	F1-Score
Improved Features	0.62	0.72	0.62	0.54
Generalized Features	0.80	0.80	0.80	0.80

Performance on the LC25000 test set.



Improved Features



Generalized Features

Feature Extraction – Discussion

- **Model 1 (Raw input)**: Retains features that PCA may discard. Higher F1 and AUC $\Rightarrow$ more balanced performance across thresholds.
- **Model 2 (PCA)**: Denoises input and highlights dominant features. Higher accuracy and recall $\Rightarrow$ fewer missed cancer cases.
- PCA improves sensitivity (recall)
- Raw features yield richer decision boundaries, better robustness to threshold changes.

Resolution – Discussion

- Evaluation on LC25000 introduces domain shift relative to LungHist700 training.
- **Model 4 (40×)** outperforms **Model 3 (20×)** across accuracy, precision, F1, and AUC.
- Recall = 1.00 for both ⇒ no false negatives.
- Reasoning: 40× better aligns with LC25000 magnification (750×750) and preserves morphological structures.

Domain Shift – Discussion

- **Within-domain (Models 1,2):** consistent performance, no distribution mismatch.
- **Cross-domain (Models 3,4):** reduced performance due to staining, scanner, and magnification differences.
- Higher resolution (Model 4) reduces but does not eliminate domain shift impact.
- Confirms that magnification consistency is a partial solution, but dataset heterogeneity remains a major limitation.

Discussion: Adenocarcinoma Differentiation

- Models capture coarse differences but fail at fine-grained subtype distinctions.
- Challenges:
  - Limited dataset size.
  - Severe class imbalance.
  - Morphological subtlety between subtypes (especially between moderate and poor differentiation).
- Potential improvements:
  - Expand dataset with additional subtype annotations (e.g., scc differentiation).
  - Incorporate texture-sensitive or nuclei-level features.
  - Stronger augmentation or regularization.

Discussion: LC25000 Generalization

- **Deep CNNs (ResNet)** transfer moderately well despite domain shift.
- **Random Forest (40×)** surprisingly robust — benefits from higher resolution texture cues.
- **Logistic Regression** underperforms; linear model and simple HOG features insufficient for transfer.
- Overall: higher magnification (40×) consistently improves cross-dataset generalization.

Multi-Scale & Feature Generalization Takeaways

- **Multi-scale helps in-distribution:** $20\times \rightarrow$ Multi-scale improves from 0.80 to 0.84 accuracy.
- **But scale alone isn't enough for OOD:** Multi-scale baseline reaches 0.66 Acc on LC25000.
- **Feature engineering trade-off:** Improved features $\uparrow$ LungHist700 (0.95 Acc) but $\downarrow$ LC25000 (0.62 Acc) due to staining specificity.
- **Generalized features $\Rightarrow$ robustness:** 0.86 (LungHist700) and 0.80 (LC25000).
- **Context vs prior work:** Prior LungHist700 results reported 90% ($20\times$) and 82% ($40\times$); our improved multi-scale RF reached 95% on LungHist700 but at the cost of OOD generalization, which the generalized features address.

References I

[1] Komura, D. & Ishikawa, S. (2018). *Machine learning methods for histopathological image analysis*. Cancers, 10(1), 1–14.

[2] Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., Cui, C., Corrado, G., Thrun, S., & Dean, J. (2019). *A guide to deep learning in healthcare*. Nature Medicine, 25(1), 24–29.

[3] Haralick, R. M., Shanmugam, K., & Dinstein, I. (1973). *Textural features for image classification*. IEEE Transactions on Systems, Man, and Cybernetics, SMC-3(6), 610–621.

[4] Ojala, T., Pietikäinen, M., & Harwood, D. (1996). *A comparative study of texture measures with classification based on featured distributions*. Pattern Recognition, 29(1), 51–59.

[5] Otsu, N. (1979). *A threshold selection method from gray-level histograms*. IEEE Transactions on Systems, Man, and Cybernetics, 9(1), 62–66.

[6] Diosdado, J., Gilabert, P., Seguí, S., & Borrego, H. (2024). *LungHist700: A dataset of histological images for deep learning in pulmonary pathology*. Scientific Data, 11, 1088.

[7] Under, R. (2020). *Lung–Colon Cancer Dataset (LC25000)*. Kaggle.

Acknowledgments

The authors gratefully acknowledge:

- **Dr. Marly Gotti**, for her mentorship during this project.
- **PRIMES**, for offering us the chance to engage in meaningful and rewarding research.
- **SWIM 2025**, for giving us the platform to share and present our work.