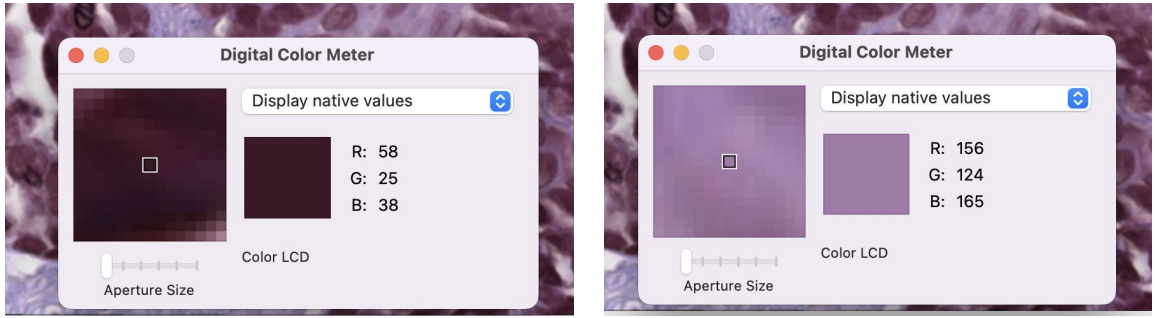
Presentation 1: Critique of Spring 2025 MIL Model

STAT 390 | Project 1 | Fall 2025

1. Color Representation - Problems

- Challenges with Current Approaches Using RGB
  - **High Correlation between RGB Channels leads to inefficient feature learning**. e.g. If a color is darker, R, G, B channels decreases together



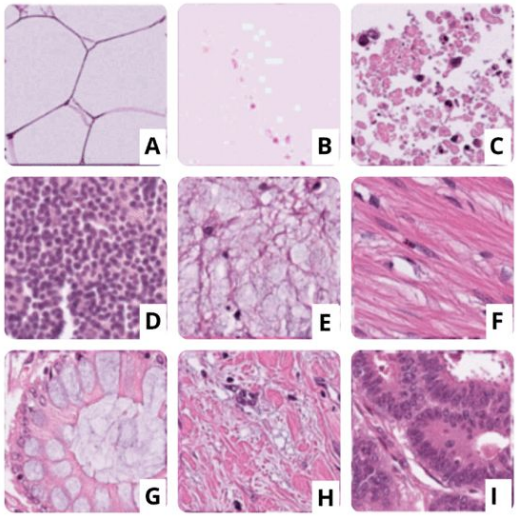
  - **Coupled Information**: RGB combines luminance (brightness/intensity of light) and chrominance (color), obscuring the structural features critical for detecting melanocyte distribution
    - Human perception relies heavily on luminance to identify boundaries and shapes. Using a color space containing luminance info could potentially enhance model interpretability

1. Color Representation - Other Color Spaces

Color Space Name	Description	Channels
YCbCr	Used in digital video and image compression, it separates luma (brightness) from two chroma (color difference) components	**Luma (Y)**, **Blue-difference Chroma (CB)**, **Red-difference Chroma (CR)**
Lab (CIELAB)	Designed to be perceptually uniform, it defines color based on lightness and two color-opponent dimensions	**Lightness (L)**, **Green-Red Axis (a)**, **Blue-Yellow Axis (b)**
HSV	Represents colors in terms of their tint, purity, and brightness, often used in color pickers	**Hue (H)**, **Saturation (S)**, **Value (V)**
XYZ (CIEXYZ)	A device-independent color space based on how an average human observer perceives color, serving as a standard reference	**Red-like (X)**, **Green-like (Y)**, **Blue-like (Z)**

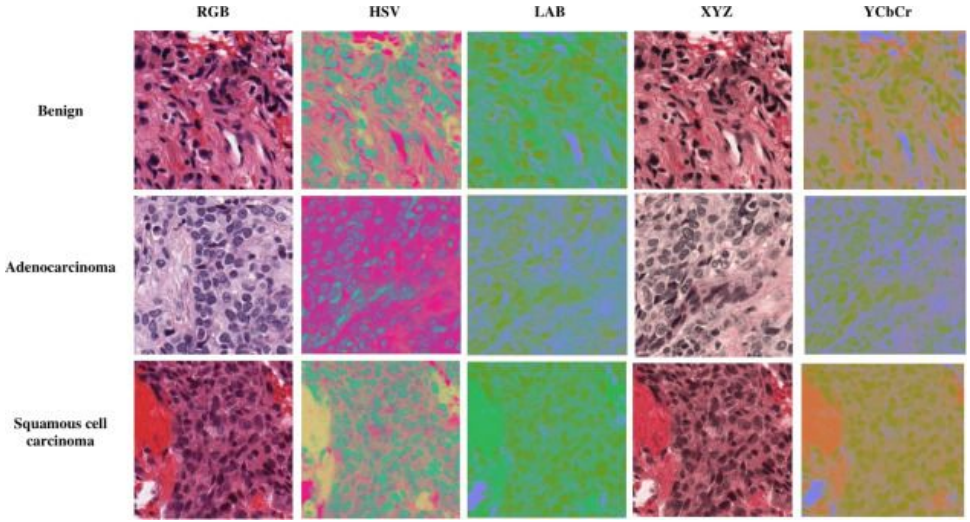
1. Color Representation - Literature Review

- Bishnoi & Goel (2023): impact of 5 color spaces on tissue slice lung cancer classification
  - Color spaces: RGB, HSV, LAB, XYZ, YCbCr
  - **HSV** consistently outperforms other color spaces in ResNet, DenseNet, and Inception V3, and their own model
- Velastegui & Pedersen (2023): 9 categories of colorectal cancer histological imaging
  - Color spaces: RGB, XYZ, HSV, YCbCr, CIELAB, LLL, AAA, BBB
  - **HSV and CIELAB** have the highest overall accuracy
  - **HSV** is the best for category D (lymphocytes) similar to our images



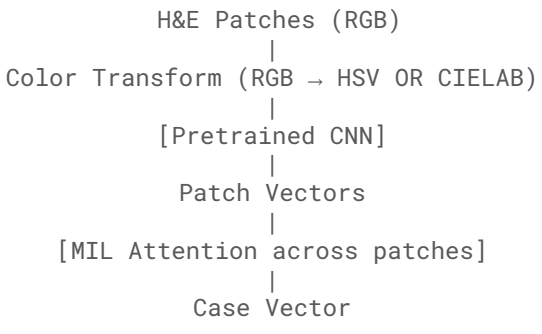
1. Color Representation - Action Plans

- Transform RGB into HSV and CIELAB and benchmark with MIL from last quarter
- Test whether channels that primarily contain brightness information (Y-channel in YCbCr, V-channel in HSV, and L-channel in CIELAB) perform well standalone
    - If so, we can consider using 3 channels across matched stains. Eg. (Y_{HE}, Y_{Mel}, Y_{Sox})

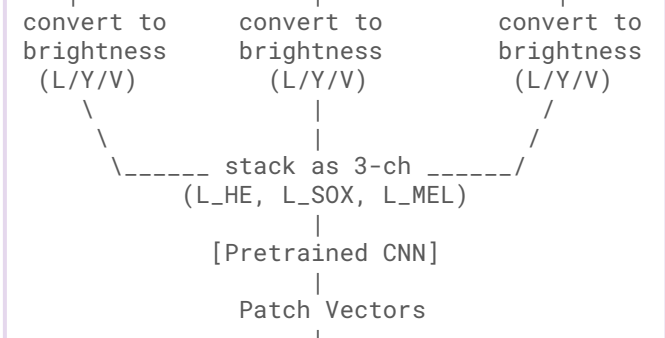


1. Color Representation - Action Plans

Option A:
Transform RGB into HSV, CIELAB



Option B:
Transform RGB into HSV and CIELAB



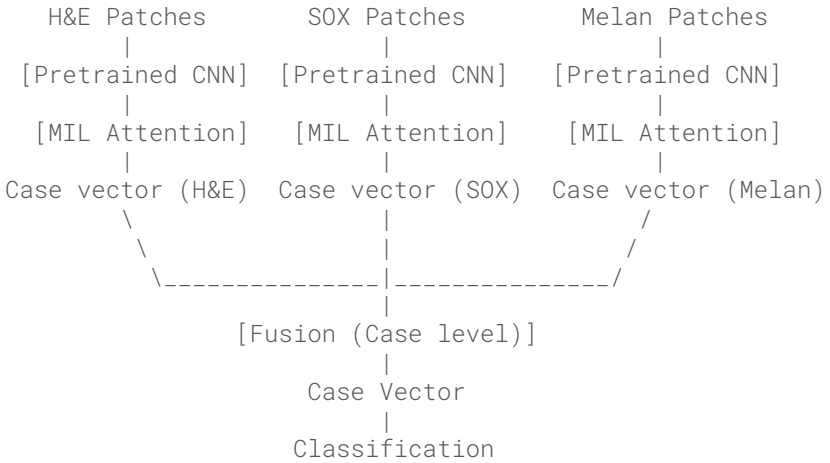
2. Multi-Stain - Overview

- **Current baseline: Prior MIL used a single stain (H&E)**
  - Risks missing complementary information available in different stains
  - Robustness issue when H&E is noisy (less serious when used case-level)

- **Our upgrade: MIL + Cross-stain**
  - Build three parallel stain branches (H&E / SOX / Melan)
  - *Two ways* to implement this
    - ***Option A:*** Each branch has its own MIL attention
      - Outputs per-stain case-level embeddings
    - ***Option B:*** Each branch does feature extraction only
      - Then, patch embeddings “fused” together; single MIL attention for final case vector

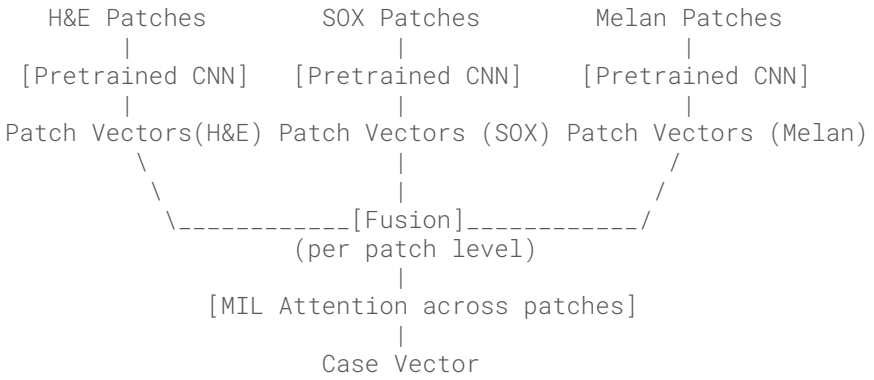
2. Multi-Stain - Architecture

Option A:
Per-stain Case Embeddings



(Three parallel MIL pipelines)

Option B:
Per-stain Patch-Level Features



(Cross-stain feature fusion; requires matching patches)

2. Multi-Stain - Comparison of Two Options

Method	Option A: Per-stain Case Embeddings	Option B: Per-stain Patch-Level Features
Implementation	MIL attention in each branch	MIL attention after per-stain feature extraction
Advantages	- Robust to patch mismatches - Robust when a stain is missing - Interpretable per-stain attention maps	- Captures cross-stain interaction for a specific cell area - Efficient: only one MIL attention module
Disadvantages	- Cross-stain interactions captured only after case summarization - Heavy (3 * MIL), potential redundancy	- Not robust to patch mismatches - Less per-stain interpretability (no per-stain attention map)

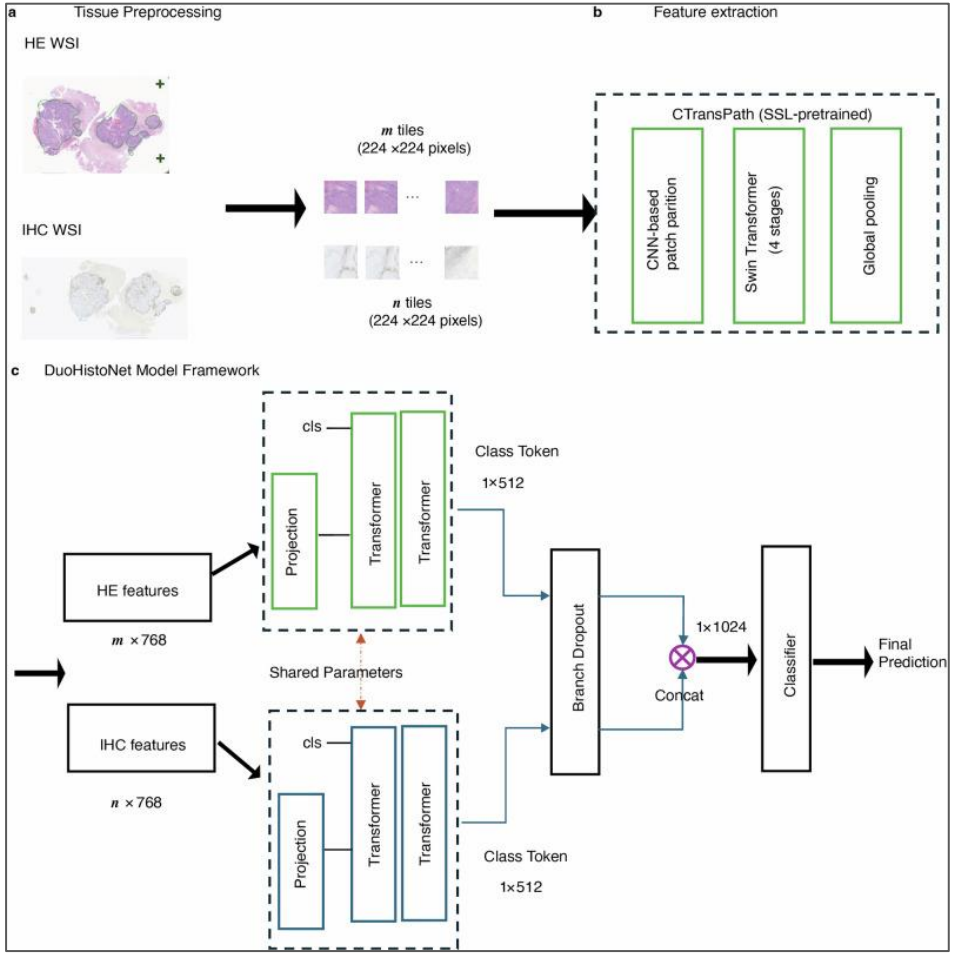
2. Multi-Stain - Literature Review

- Foersch et al., Nature Medicine (2023) *Link*
  - Built a multi-stain deep learning model using H&E + multiple IHC stains for colorectal cancer.
  - Does not implement MIL attention; uses gated attention instead
    - Otherwise similar to *Option B:* patch-level feature extraction and fusion

  - Multi-stain architecture improves…
    - Accuracy: better accuracy than using one single stain
    - Interpretability: attention weights aligned with known biology (e.g., high weights for certain cell features).

2. Multi-Stain - Literature Review

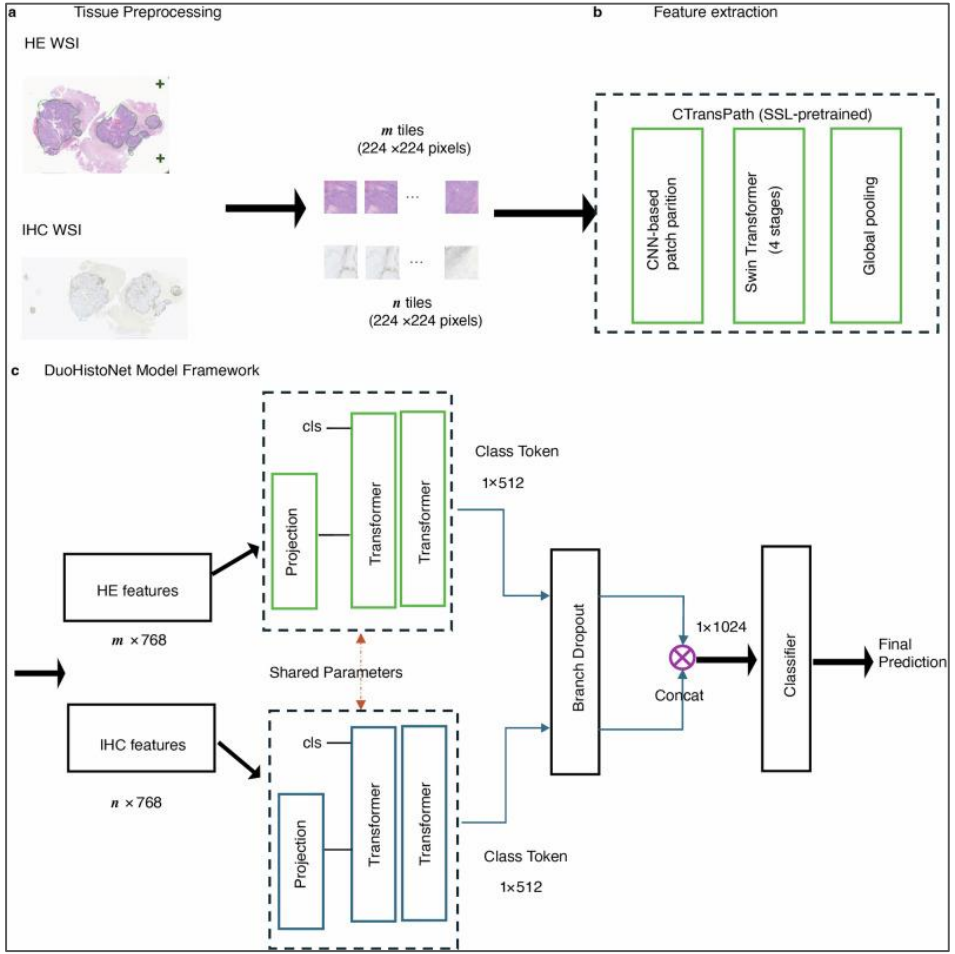
- Cheng et al., Communications Medicine (2025) (DuoHistoNet) *Link*
    - **Dual-branch, per-stain transformer**
        - patch-level feature extraction and attention-based MIL aggregation
        - Each branch outputs a case-level embedding
        - Resembles *Option A*
    - Outputs from the two branches are concatenated and fed into a case-level classifier
    - Result: Significantly outperformed single-stain models



2. Multi-Stain - Literature Review

Cheng et al., Communications Medicine (2025) (DuoHistoNet) [Link]

- **Other highlights of this model**
  - Transformer aggregator (MHSA): authors of this article claim that this is superior to other MIL approaches
  - Each tile is passed through a CNN + Swin Transformer hybrid base model first
    - 768-dim feature embedding
  - Then, per-stain MIL



2. Multi-Stain - Discussion

- Should we preserve stain-specific MIL attention (option A), or should we apply MIL attention on cross-stain feature embeddings?
- Do we expect synergy mainly at the patch level (feature level) or the case level?

3. Image Size – Current Approaches

Adaptive Pooling (without resize)

- Use ResNet built-in adaptive pooling layer with batch size = 4
  - Only worked well for Sox10 stain (recall of 0.88)
- Increasing batch size from 4 to 8 decreases validation performance volatility and improves performance

Resizing

- Images are resized to 224x224 and can distort information
- Use built-in ResNet adaptive pooling layer
  - Consistent results for all 3 stains, but Sox10 still the strongest
- Resizing + adaptive pooling + increased batch size shows most promising results

Source: Veer's Spring 2025 Findings

3. Image Size – Proposed Solutions

Areas to explore

- Continue testing additional combinations of image resizing methods
  - adaptive pooling, resizing, cropping, random horizontal flip

- Test higher batch sizes fed into the model given computational restraints
  - Ex. batch size = 16 or 32 for KimiaNet and ResNet

- Minimize padding (currently number buckets = batch size)
  - Higher number of buckets and lower batch size had best results (David and Veer)
  - Ideas: Number of buckets follow distribution of image sizes, find optimal buckets given computational restraints

4. Training, Testing, & Validation Sets

Problems

- Training set might not have all categories / patterns of high-grade patches
- We don't have a categorization of high-grade patterns to refer to in stratifying the train set

Proposed Solutions

- Calculate variation in accuracy in under-performing cases when different cases are in the training set
- Develop categorization to ensure all high-grade patterns are present in training set

4. Training, Testing, & Validation Sets

Proposed method:

Step 1: Identify which cases underperform (are misclassified often)
→ Run experiment using ~8-fold CV and 2-3 different seeds to identify which cases' classification success vary by training set

Step 2: Extract patch embeddings from model and cluster high-grade patches to find distinct pattern types
→ Use k-means or HDBSCAN for clustering

Step 3: Identify which clusters patches in underperforming cases fall into

Step 4: Oversample or add similar cases in training set accordingly

References

Madusanka, N., Jayalath, P., Fernando, D., Yasakethu, L., & Lee, B.-I. (2023). Impact of H&E Stain Normalization on Deep Learning Models in Cancer Image Classification: Performance, Complexity, and Trade-Offs. Cancers, 15(16), 4144. https://doi.org/10.3390/cancers15164144

Bishnoi, Vidhi, and Nidhi, Goel. (2023). A Color-Based Deep-Learning Approach for Tissue Slide Lung Cancer Classification. Biomedical Signal Processing and Control. www.sciencedirect.com/science/article/abs/pii/S1746809423005840

Questions for Krish on Tuesday

- Did the MIL group use the resized images? If so, why? (does their model mandate that?) Update: because they used densenet. Did DeIT or Kimeanet seem as promising? *Yes, resized to 224*224*
- Total presentation time *"Time doesn't matter"*
- Can we get a categorization from doctors? Or get them to vibe check our unsupervised clustering results
  - *He will ask (i don't think we can get the categories from the doctors)*
  - *He suggested that we test whether changing the training set, for example by using a different combination of cases, would affect the model's performance, particularly the accuracy on a specific case.*
    - *For example, if model performance on case #4 (in the test set) is poor, see if changing the combination of the training set will boost/worsen the performance on case #4*
  - *Unsupervised clustering might be a good approach, but what it does is finding the difference between cases, rather than finding what features/patterns the model struggles with*