5/21 Presentation

Hard to Classify Cases

Benign

- Case 34: (0.17, 0.58, 0.47)
- Case 36: (0.43, 0.57, 0.42)
- Case 46: (0.17, 0.04)
- Case 53: (0.28)
- Case 85: (0.35, 0.17)
- Case 86: (0.33)

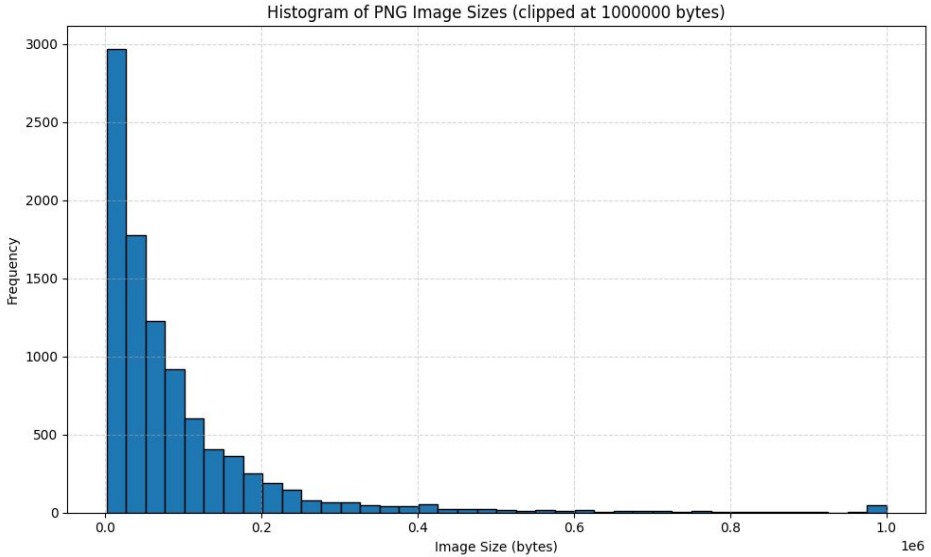
High Grade

- Case 44: (0.34, 0.51)
- Case 87: (0.33)

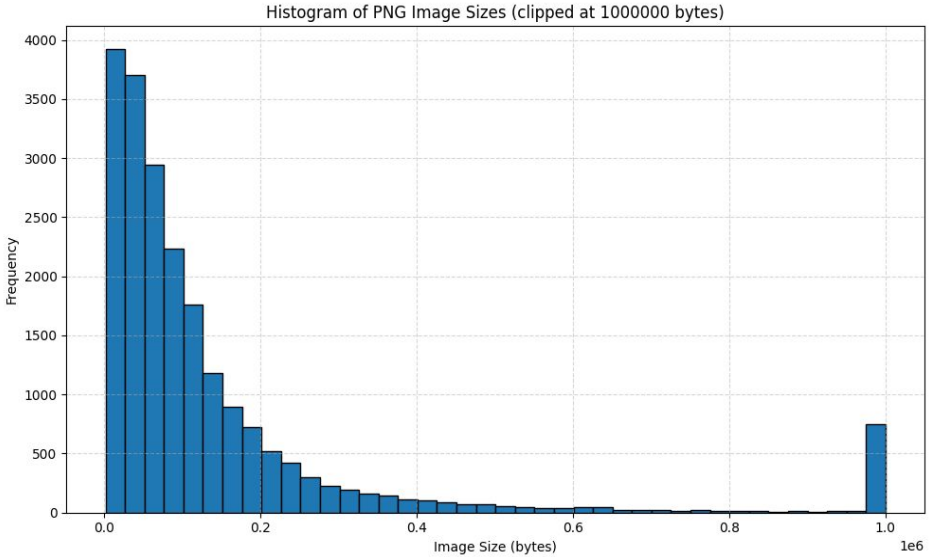
Why did new patches hurt performance on all H&E cases vs just the good cases?

Old vs New Image Size Distribution

OLD



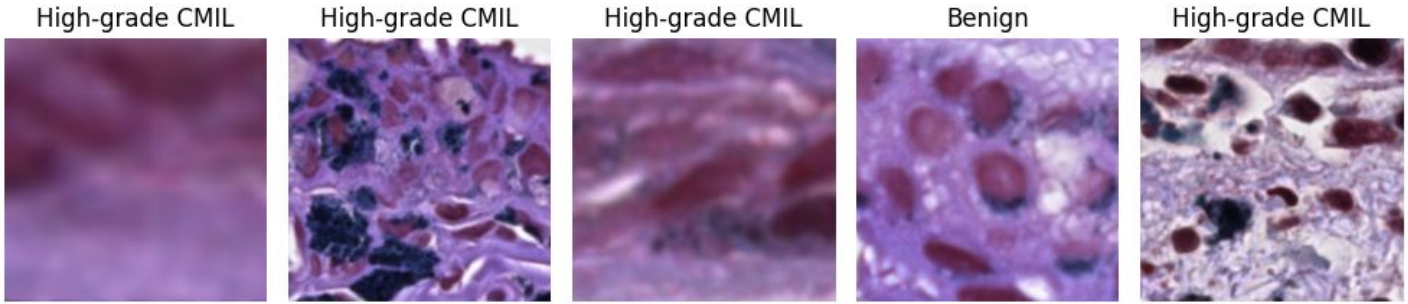
NEW



Old vs New Data Splits

📊 TRAIN SET SUMMARY:
Total cases: 28
Total patches: 4686
  Benign patches: 1564
  High-grade patches: 3122
📊 VAL SET SUMMARY:
Total cases: 8
Total patches: 1340
  Benign patches: 385
  High-grade patches: 955
📊 TEST SET SUMMARY:
Total cases: 9
Total patches: 1507
  Benign patches: 479
  High-grade patches: 1028

OLD



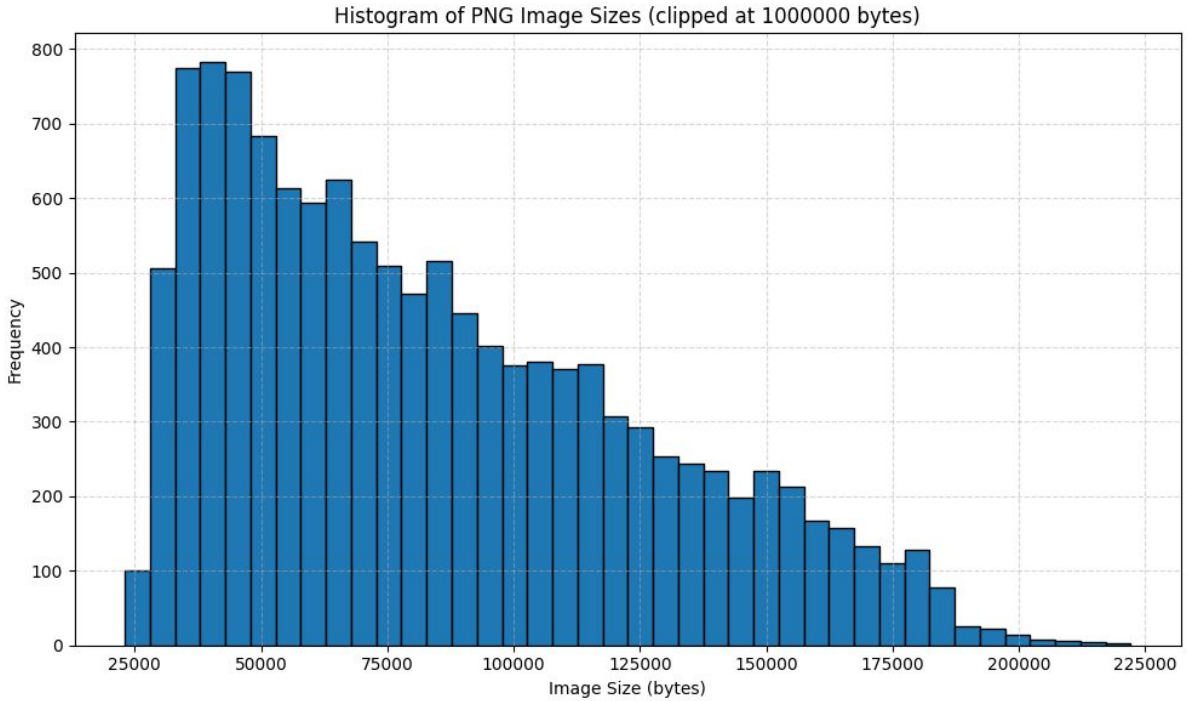
📊 TRAIN SET SUMMARY:
Total cases: 40
Total patches: 12229
  Benign patches: 3770
  High-grade patches: 8459

📊 VAL SET SUMMARY:
Total cases: 11
Total patches: 3292
  Benign patches: 736
  High-grade patches: 2556
📊 TEST SET SUMMARY:
Total cases: 13
Total patches: 4302
  Benign patches: 1602
  High-grade patches: 2700

NEW

Middle 60%

📊 TRAIN SET SUMMARY:
Total cases: 38
  Total patches: 7457
  Benign patches: 1927
  High-grade patches: 5530
📊 VAL SET SUMMARY:
Total cases: 10
Total patches: 2309
  Benign patches: 1152
  High-grade patches: 1157
📊 TEST SET SUMMARY:
Total cases: 13
  Total patches: 2137
  Benign patches: 939
  High-grade patches: 1198



Comparison Total vs Middle 60%

Res

	precision	recall	f1-score	support
0	0.31	0.34	0.32	1602
1	0.63	0.60	0.62	3062
accuracy			0.51	4664
macro avg	0.47	0.47	0.47	4664
weighted avg	0.52	0.51	0.52	4664

	precision	recall	f1-score	support
0	0.53	0.17	0.26	939
1	0.57	0.88	0.70	1198
accuracy			0.57	2137
macro avg	0.55	0.52	0.48	2137
weighted avg	0.55	0.57	0.50	2137

CNN

	precision	recall	f1-score	support
Benign	0.06	0.02	0.03	1713
High-grade CMIL	0.42	0.66	0.52	1849
accuracy			0.35	3562
macro avg	0.24	0.34	0.27	3562
weighted avg	0.25	0.35	0.28	3562

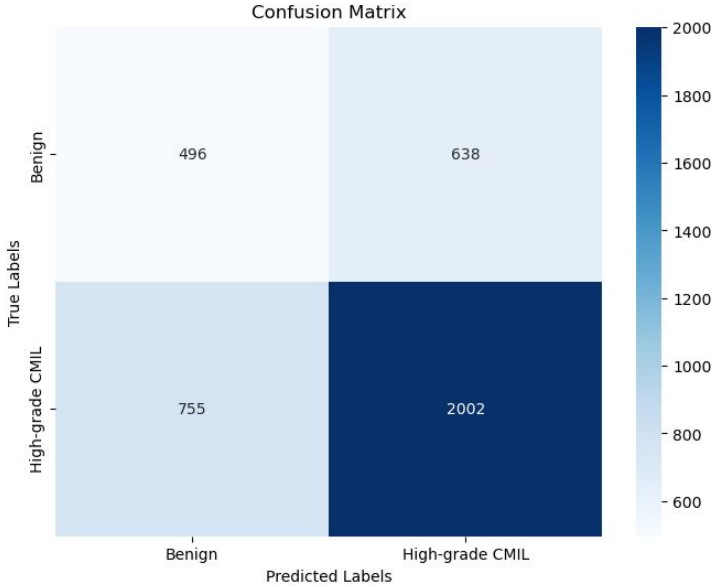
	precision	recall	f1-score	support
Benign	0.35	0.17	0.23	939
High-grade CMIL	0.54	0.76	0.63	1198
accuracy			0.50	2137
macro avg	0.44	0.46	0.43	2137
weighted avg	0.45	0.50	0.45	2137

CoAtNet

	precision	recall	f1-score	support
Benign	0.36	0.24	0.29	1602
High-grade CMIL	0.62	0.75	0.68	2700
accuracy			0.56	4302
macro avg	0.49	0.49	0.48	4302
weighted avg	0.53	0.56	0.54	4302

	precision	recall	f1-score	support
Benign	0.64	0.21	0.32	939
High-grade CMIL	0.59	0.90	0.72	1198
accuracy			0.60	2137
macro avg	0.62	0.56	0.52	2137
weighted avg	0.61	0.60	0.54	2137

All H&E patches > 60 pixels (ResNet with CBAM):



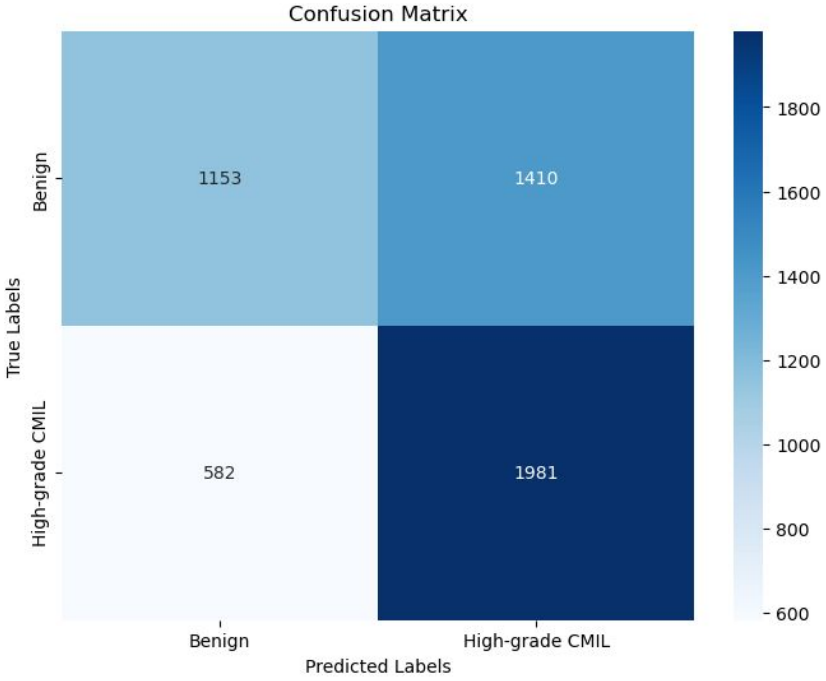
Classification Report:

	precision	recall	f1-score	support
Benign	0.40	0.44	0.42	1134
Malignant	0.76	0.73	0.74	2757
accuracy			0.64	3891
macro avg	0.58	0.58	0.58	3891
weighted avg	0.65	0.64	0.65	3891

Resampling

Method: oversampling

- Duplicating patches from the minority class in H&E (benign) until classes are balanced, because we don’t want to lose a ton of high-grade patches
- Slightly lower accuracy than before
- Hypothesis: we are overfitting now
- Next steps: change weights of benign class instead of oversampling

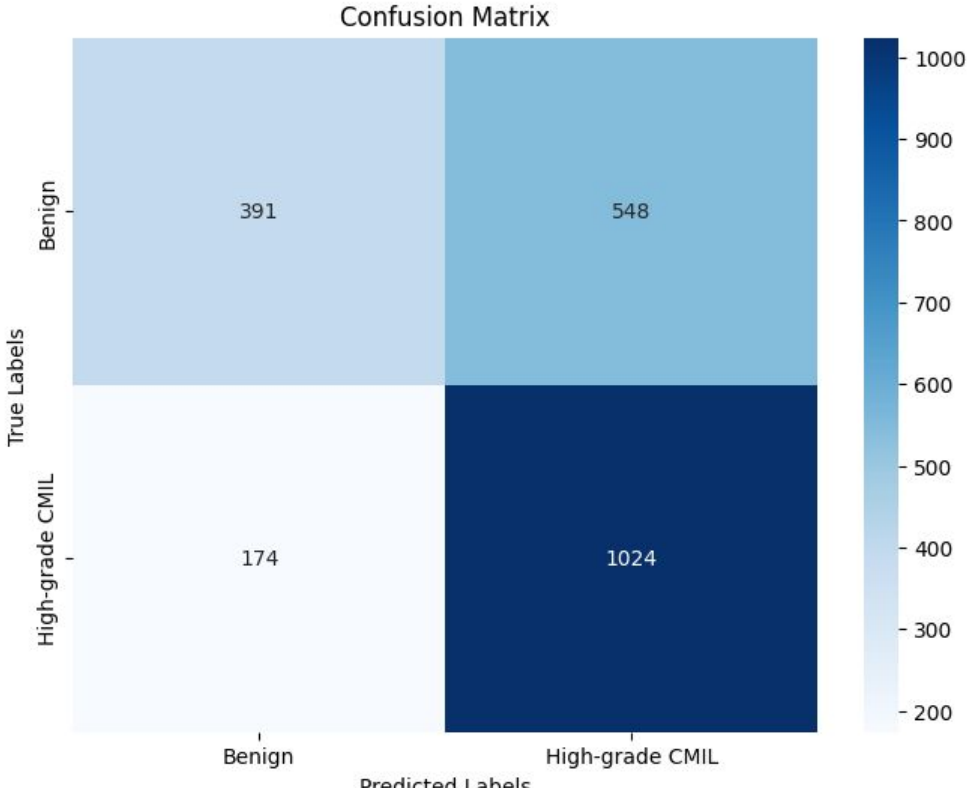



```
Accuracy:  0.6114
Precision: 0.5842
Recall:    0.7729
F1 Score:  0.6654
```

DeiT

	precision	recall	f1-score	support
0	0.69	0.42	0.52	939
1	0.65	0.85	0.74	1198
accuracy			0.66	2137
macro avg	0.67	0.64	0.63	2137
weighted avg	0.67	0.66	0.64	2137

```
Case 95: Accuracy = 0.9796, True Label = High-grade CMIL
Case 86: Accuracy = 0.2299, True Label = Benign
Case 57: Accuracy = 0.9858, True Label = High-grade CMIL
Case 4: Accuracy = 0.4384, True Label = Benign
Case 67: Accuracy = 1.0000, True Label = High-grade CMIL
Case 72: Accuracy = 0.8276, True Label = High-grade CMIL
Case 99: Accuracy = 0.9600, True Label = High-grade CMIL
Case 2: Accuracy = 0.5543, True Label = Benign
Case 7: Accuracy = 0.8193, True Label = High-grade CMIL
Case 64: Accuracy = 0.9773, True Label = High-grade CMIL
Case 87: Accuracy = 0.6087, True Label = High-grade CMIL
Case 38: Accuracy = 0.4016, True Label = High-grade CMIL
Case 56: Accuracy = 0.1481, True Label = Benign
```



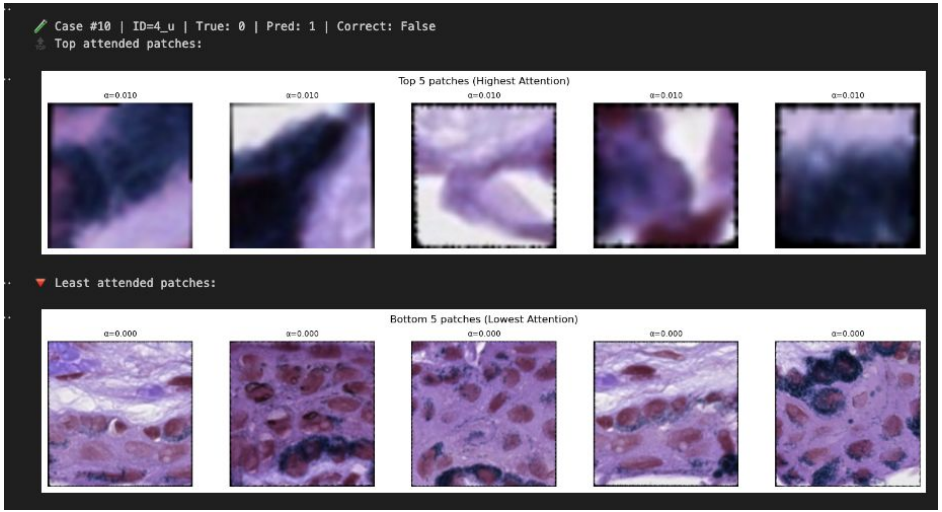
Additionally Epochs Needed


```
Epoch: 1/5.. Training Loss: 0.322.. Validation Loss: 0.710.. Validation Accuracy: 0.773
✅ Checkpoint saved: /content/drive/MyDrive/checkpoints/20250521_085223_coatnet_0_epoch1.pth
```

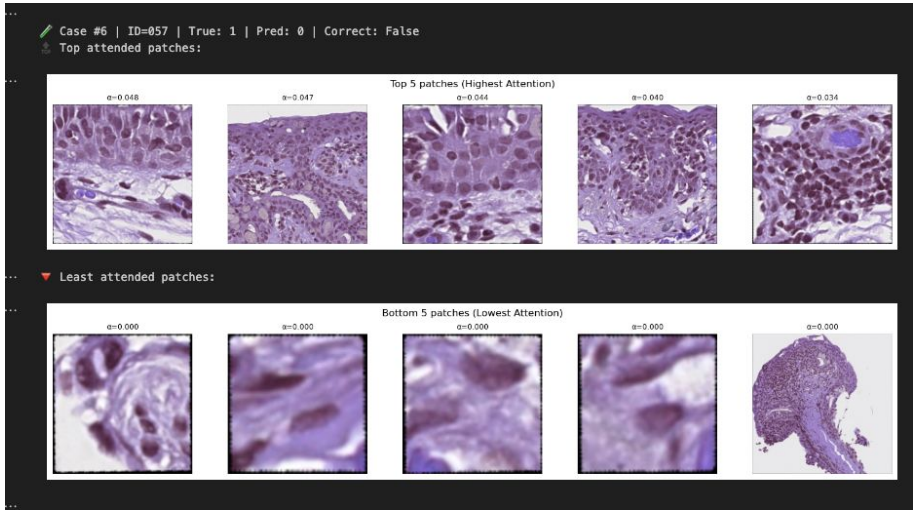
MIL- Visualizing Attention on Patches

why does blurry / non-informative patches have high weight?

Densenet w new patches (65%)

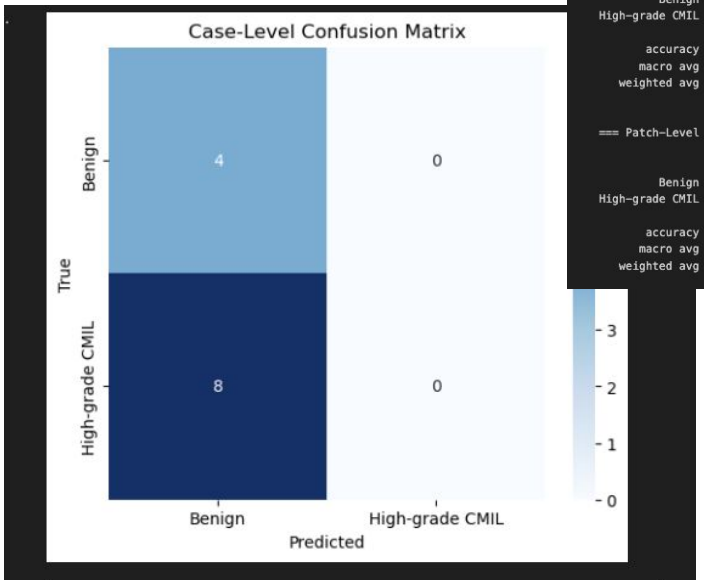


Densenet w curated patches (73%)



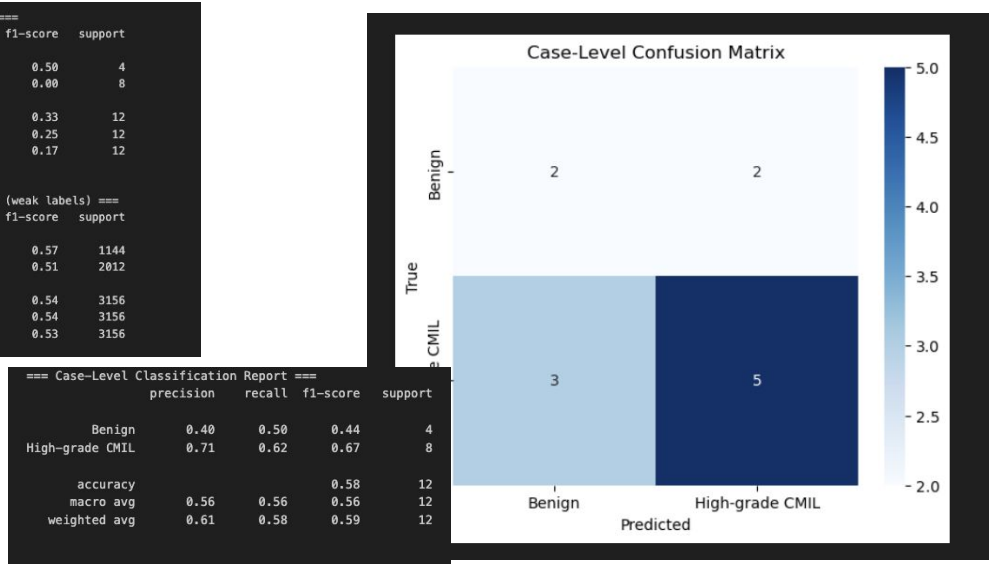
Blurry patches got high attention in MIL likely because the patch classifier used didn't learn strong discriminative features. With a better patch classifier (trained on curated, high-quality data), MIL focuses attention on more meaningful regions.

Accuracy for MIL w Densenet using curated patches



MIL w densenet on curated patches

```
=== Case-Level Classification Report ===
                precision    recall  f1-score   support

        Benign       0.33      1.00      0.50         4
High-grade CMIL      0.00      0.00      0.00         8

      accuracy                           0.33        12
     macro avg       0.17      0.50      0.25        12
  weighted avg       0.11      0.33      0.17        12


=== Patch-Level Classification Report (weak labels) ===
                precision    recall  f1-score   support

        Benign       0.43      0.84      0.57      1144
High-grade CMIL      0.80      0.37      0.51      2012

      accuracy                           0.54      3156
     macro avg       0.62      0.61      0.54      3156
  weighted avg       0.67      0.54      0.53      3156
```



MIL w new patch densenet

```
=== Case-Level Classification Report ===
                precision    recall  f1-score   support

        Benign       0.40      0.50      0.44         4
High-grade CMIL      0.71      0.62      0.67         8

      accuracy                           0.58        12
     macro avg       0.56      0.56      0.56        12
  weighted avg       0.61      0.58      0.59        12


=== Patch-Level Classification Report (weak labels) ===
                precision    recall  f1-score   support

        Benign       0.18      0.14      0.15      1831
High-grade CMIL      0.57      0.64      0.60      3250

      accuracy                           0.46      5081
     macro avg       0.37      0.39      0.38      5081
  weighted avg       0.43      0.46      0.44      5081
```

Cross-stain patch dataset

- 22 cases, 12 high-grade and 10 benign
- 13242 individual patches matched into 4414 triplets
- 75-25 train/test split
- Both train and test sets are 70% high grade, 30% benign at patch level
- 20-80 percentile pixel size filter as per David’s code
- Averaging patches within triplets that are <90% blank pixels

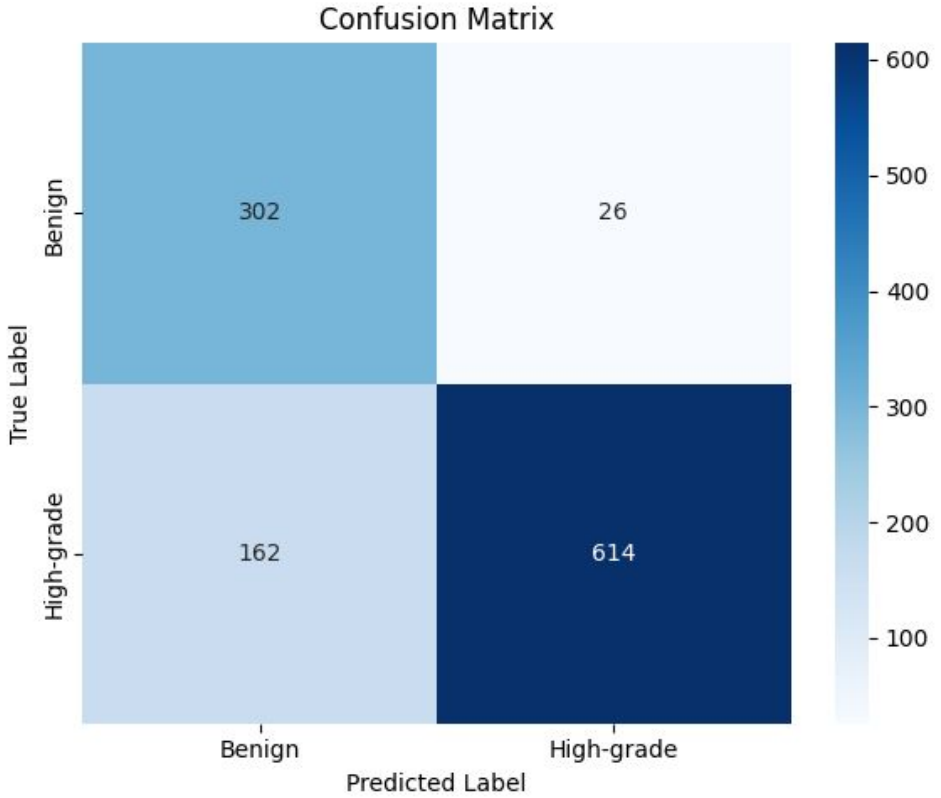
David’s Alexnet pipeline with cross-stain patches

Cross stain patches

```
Test Patch-level Accuracy : 82.97%
Precision                 : 0.9594
Recall                    : 0.7912
F1 Score                  : 0.8672
```

Regular patches

```
Test Patch-level Accuracy : 90.40%
Precision                 : 0.8596
Recall                    : 0.8596
F1 Score                  : 0.8596
```



Cross-stain Alexnet pipeline compiled with focal loss

```
Test Patch-level Accuracy : 87.95%
Precision                 : 0.9181
Recall                    : 0.9098
F1 Score                  : 0.9139
```

Weighted loss with tunable parameters gamma and alpha

Gamma: values > 0 place more weight on hard-to-learn examples

Alpha: weights for the positive and negative classes (1 - alpha weight for negative class) to address imbalance