# Beyond Benchmark Accuracy: Cross-Dataset Generalization and Open-Set Safety in Gastrointestinal Endoscopy AI

**Mesut Erkin Ozokutgen, Adebayo Adewumi Abayomi-Alli, Sema Nur Alan**

Vilnius Gediminas Technical University (Vilnius Tech), Vilnius, Lithuania | adebayo-adewumi.abayomi-alli@vilniustech.lt

## 1 Why external validation matters

- Benchmark accuracy is measured on test images from the **same source** as the training data; clinical use brings other devices, other patients and **findings the model has never seen**.
- We test **external generalization**, **confidence calibration** and **open-set behavior** of GI endoscopy classifiers.

## 2 Data and label harmonization

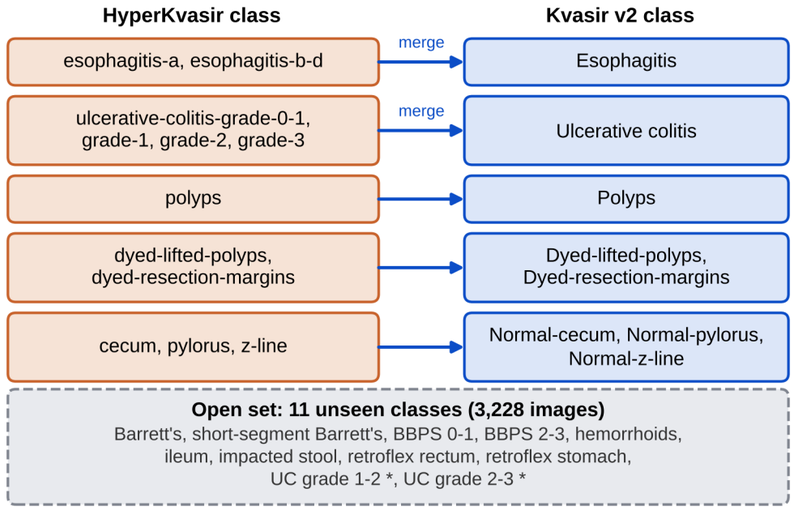



**Figure 1.** HyperKvasir Label Harmonization and Open-Set Classes..

## 3 Experimental design

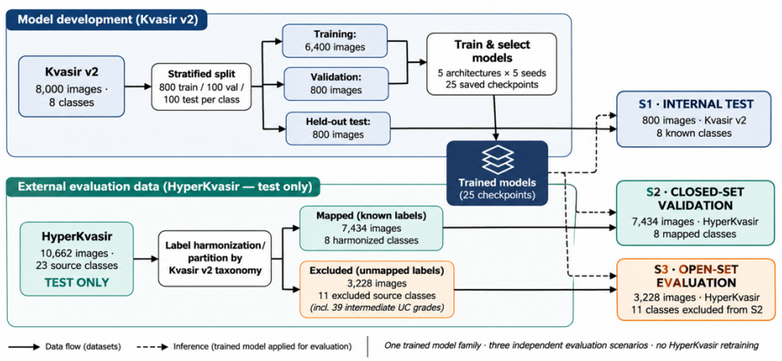



**Figure 2.** Pipeline. Kvasir v2 split 800/100/100 per class; HyperKvasir is test only.

- **Models:** EfficientNet-B1, EfficientNet-B3, ConvNeXt-Tiny, DenseNet121, Swin-T; 5 seeds each (42, 55, 66, 123, 423).
- **Setup:** 224 x 224 input, no augmentation, no post-hoc calibration.
- **Metrics:** accuracy, Macro-F1, AUROC, expected calibration error (ECE); abstention at confidence < 0.8. Mean ± SD over seeds.

## 4 Closed-set results: internal vs external

**Table 1.** Mean ± SD over 5 seeds. External: 7,434 HyperKvasir images (8 shared classes). Internal: 800 Kvasir v2 test images.

| Model | Internal Macro-F1 | External accuracy (%) | External Macro-F1 | External ECE |
|---|---|---|---|---|
| EfficientNet-B1 | 0.936 ± 0.003 | 96.22 ± 0.42 | 0.959 ± 0.005 | 0.016 ± 0.007 |
| EfficientNet-B3 | 0.928 ± 0.009 | 95.97 ± 0.41 | 0.956 ± 0.005 | 0.018 ± 0.008 |
| **ConvNeXt-Tiny** | 0.932 ± 0.009 | 96.56 ± 0.14 | **0.963 ± 0.002** | 0.023 ± 0.004 |
| DenseNet121 | 0.928 ± 0.003 | 95.76 ± 0.36 | 0.954 ± 0.004 | 0.014 ± 0.004 |
| Swin-T | 0.918 ± 0.010 | 94.08 ± 0.63 | 0.935 ± 0.008 | 0.021 ± 0.010 |

- **ConvNeXt-Tiny** was best externally: 96.6% accuracy, Macro-F1 0.963, AUROC 0.997.
- External Macro-F1 (0.935 to 0.963) was **at least as high as internal** (0.918 to 0.936) for every model.
- Calibration stayed good: external ECE 0.014 to 0.023.

## 5 Class-level view

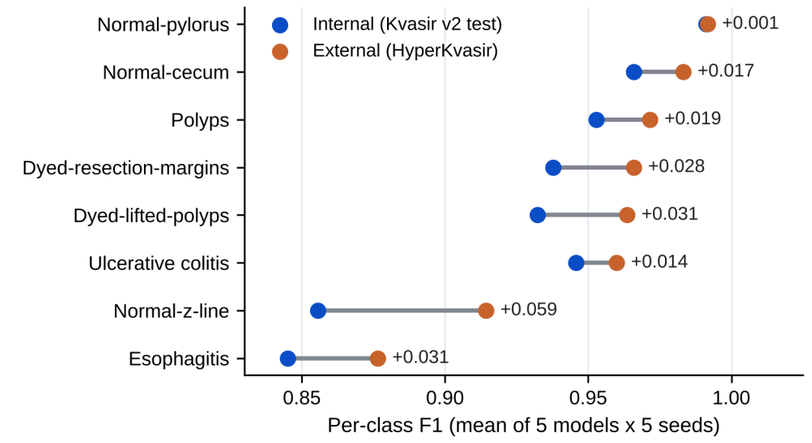
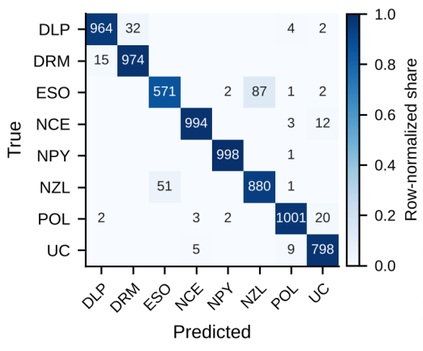



**Figure 3.** Per-class F1, internal vs external (change shown at right).




- **Esophagitis** is the weakest class (external F1 0.83 to 0.90); on average **12%** of its images are predicted as **normal z-line**.

**Figure 4.** External confusion matrix, EfficientNet-B1 (seed 423). DLP dyed-lifted polyps, DRM dyed resection margins, ESO esophagitis, NCE normal cecum, NPY normal pylorus, NZL normal z-line, POL polyps, UC ulcerative colitis.

## 6 Open-set behavior on unseen findings

**Table 2.** Pooled over 25 runs. Share = non-abstained predictions sent to that class; for hemorrhoids and retroflex stomach the class varied between runs.

| Unseen class (n) | Mostly predicted as | Share (%) | Mean conf. | Rejected (%) |
|---|---|---|---|---|
| Barrett's (94) | esophagitis | 88.45 | 0.91 | 17.32 |
| Ileum (9) | ulcerative-colitis | 96.48 | 0.94 | 11.11 |
| Hemorrhoids (6) | ulcerative-colitis | 73.67 | 0.75 | 51.33 |
| Retroflex stomach (764) | normal-pylorus | 56.09 | 0.71 | 58.85 |

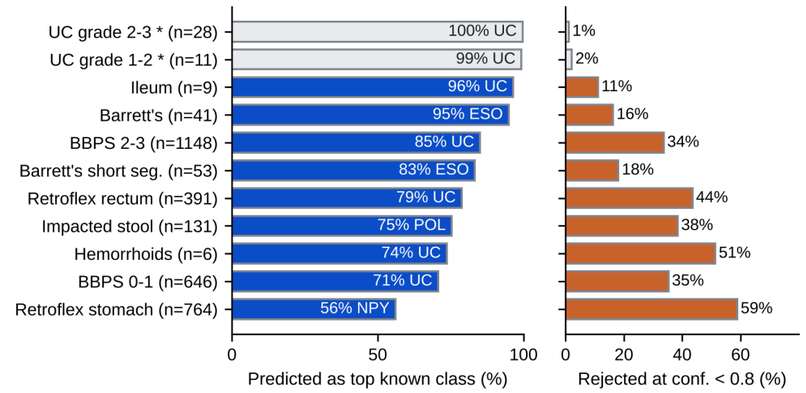



**Figure 5.** Open-set prediction patterns (means over 25 runs). ESO esophagitis, UC ulcerative colitis, POL polyps, NPY normal pylorus. * Intermediate UC grades.

- Unseen findings go to **clinically related known classes** with high confidence: Barrett's to esophagitis (88.45%, confidence 0.91), ileum to ulcerative colitis (96.48%, 0.94).
- Abstention at 0.8 rejected only **17.32%** of Barrett's and **11.11%** of ileum images, versus 58.85% of retroflex-stomach images.

## 7 Conclusions

- Strong benchmark and cross-dataset accuracy **does not establish clinical readiness**.
- Class-level and calibration checks expose weak spots, such as esophagitis, that averages hide.
- A confidence threshold alone **does not stop overconfident errors** on unfamiliar findings.

## 8 Limitations and next steps

- Both datasets come from one hospital (Bærum Hospital, Norway): a **mild shift**; image overlap not ruled out.
- Merged labels lose detail; hemorrhoids (6) and ileum (9) images are few; no significance tests.
- Next: explainability, duplicate-image check, reliability diagrams, class-by-class open-set analysis.