Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Beyond Binary: Four-Class Risk Stratification from Gastrointestinal Endoscopy Using Asymmetric-Cost Lightweight CNN–Transformer Learning

Azizur Rahman

azizurusa22@gmail.com

Indiana Wesleyan University https://orcid.org/0009-0007-6945-6953

Nakib Uddin Ahmed

nakibuddin33@gmail.com

Shahjalal University of Science and Technology https://orcid.org/0009-0006-6515-5750

Kadirur Rahman Chowdhury

Birmingham City University

Mehjabin Ferdous

St. Francis College

Md. Shahadat Hossain

Metropolitan University

Md. Amir Hossain

Metropolitan University

Research Article

Keywords: Gastrointestinal endoscopy, Risk stratification, Asymmetric loss function, Lightweight neural networks, Vision Transformer

Posted Date: July 28th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10475241/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Beyond Binary: Four-Class Risk Stratification from Gastrointestinal Endoscopy Using Asymmetric-Cost Lightweight CNN–Transformer Learning

Azizur Rahman[1] Nakib Uddin Ahmed[2] Kadirur Rahman Chowdhury[3]
Mehjabin Ferdous[4] Md. Shahadat Hossain[5] Md. Amir Hossain[5]

[1]Department of Computer Technology (CAPS), Indiana Wesleyan University, Marion, IN 46953, USA. Correspondence: azizurusa22@gmail.com. ORCID: 0009-0007-6945-6953

[2]Department of Information Technology, Shahjalal University of Science and Technology, University Avenue, Sylhet-3114, Bangladesh. nakibuddin33@gmail.com. ORCID: 0009-0006-6515-5750

[3]Birmingham City Business School, Faculty of Business, Law and Social Sciences, Birmingham City University, Birmingham B4 7BD, UK. kadirurrahmanchowdhury@gmail.com

[4]St. Francis College, 179 Livingston St, Brooklyn, NY 11201, USA. mehjabinferdousmu@gmail.com

[5]Department of Computer Science and Engineering, Metropolitan University, Bateshwar, Sylhet-3104, Bangladesh. shahadat.hossain.bio@gmail.com; amirhossain.cse113@gmail.com

Abstract

Gastrointestinal cancers account for more than 3.5 million deaths annually, yet existing AI-assisted endoscopy systems reduce the clinical decision to a binary lesion-present/absent output misaligned with published clinical practice guidelines. We present a lightweight CNN–Transformer benchmark for four-class gastrointestinal lesion *risk stratification* aligned with American College of Gastroenterology (ACG) and European Society of Gastrointestinal Endoscopy (ESGE) guidelines. We mapped 21 of 23 HyperKvasir finding classes to four clinically actionable risk

tiers—Normal, Inflammatory, Pre-malignant, and High-Risk—and trained three lightweight architectures (DenseNet-121, EfficientNet-B0, DeiT-Tiny; all <8 M parameters) under a novel **Asymmetric Endoscopy Loss (AEL)** that assigns a 5× misclassification penalty to High-Risk lesions. CLAHE preprocessing, WeightedRandomSampler, and MC Dropout uncertainty quantification ($T = 30$ passes, $\tau = 0.75$) complete the pipeline. DenseNet-121 achieved macro F1 = 0.84 on HyperKvasir; EfficientNet-B0 leads zero-shot generalization at 0.78 on the independent Kvasir-v2 cohort. Zero High-Risk lesions were missed across all architectures under the MC Dropout referral protocol—a guarantee enforced by mandatory escalation of all Pre-malignant and High-Risk predictions—with 44.9 % of cases autocleared without endoscopist review. All models produced confidence estimates in the range ECE = 0.08–0.09. GradCAM saliency maps confirm that risk-tier decisions are grounded in clinically relevant mucosal features.

Keywords: Gastrointestinal endoscopy; Risk stratification; Asymmetric loss function; Lightweight neural networks; Vision Transformer

1 Introduction

Gastrointestinal cancers—encompassing colorectal, gastric, and esophageal malignancies—collectively account for more than 3.5 million deaths annually and represent the second leading cause of cancer mortality worldwide [15]. Early-stage lesions are missed in up to 26 % of endoscopic procedures due to subtle mucosal changes, inter-endoscopist variability, and endoscopist fatigue during high-volume screening sessions [7]. Computer-aided detection (CAD) systems have demonstrated statistically significant improvements in polyp detection in randomized controlled trials [19], leading to regulatory approval of the first AI endoscopy devices. However, the clinical utility of deployed CAD systems remains constrained by a fundamental design mismatch: virtually all systems reduce the endoscopic decision to a binary lesion-present or lesion-absent output [9].

This binary framing does not correspond to any published clinical practice guideline. Both the ACG and the ESGE classify GI findings into tiered risk categories that map to distinct clinical actions: routine surveillance intervals for normal mucosa, medical management for inflammatory lesions, mandatory biopsy and intensified surveillance for pre-malignant findings, and immediate endoscopic resection or oncology referral for high-risk lesions [4, 14]. A binary AI output cannot guide these decisions, reducing the

technology to a simple alert with no actionable resolution. Binary systems also apply identical classification costs to all errors, ignoring the critical asymmetry in clinical consequences: a missed High-Risk lesion may progress to invasive cancer within months, while over-triaging a Normal finding results only in an unnecessary repeat procedure [11].

Prior deep learning work on GI endoscopy has focused predominantly on polyp detection [16] and binary lesion classification [9]. Multi-class staging has been explored in isolated contexts—such as gastric subepithelial lesion risk grading [3]—but the full GI tract has not been framed as a unified four-class ACG/ESGE-aligned risk stratification benchmark.

We address this gap through three interconnected design decisions. Architecturally, compact Vision Transformers—specifically DeiT-Tiny [18] (5.9 M parameters)—enable global mucosal context modeling through patch-token self-attention without the 86 M+ parameter overhead of full ViT [2], making them deployable on clinical-grade hardware alongside established convolutional networks. For uncertainty quantification, Monte Carlo Dropout [5] approximates Bayesian posterior inference by retaining dropout at inference time; we extend its use with a hard safety constraint that escalates all Pre-malignant and High-Risk predictions regardless of model confidence, encoding clinical risk asymmetry directly into the referral logic. For the loss function, Focal Loss [8] addresses class imbalance by down-weighting easy examples but does not encode inter-class *risk* asymmetry—the clinically critical distinction between missing a Normal finding and missing a High-Risk lesion. Cost-sensitive learning has been applied in medical image classification where class imbalance and asymmetric misclassification costs coexist, but explicit survival-weighted cross-entropy formulations for multi-tier GI risk stratification have not previously been reported.

We reframe the problem as *four-class risk stratification* aligned with published ACG/ESGE intervention thresholds rather than binary lesion presence. The core technical contribution is the **Asymmetric Endoscopy Loss (AEL)**, a weighted cross-entropy formulation that assigns a 5× penalty to High-Risk misclassification, encoding the survival cost asymmetry directly into the training objective. Three lightweight architectures (<8 M parameters each: DenseNet-121, EfficientNet-B0, DeiT-Tiny) are benchmarked under identical conditions and evaluated zero-shot on the independent Kvasir-v2 cohort. Monte Carlo Dropout uncertainty quantification constructs a referral

protocol under which zero High-Risk lesions were missed on the internal test set.

GradCAM saliency maps provide per-class explainability for clinical audit; an ablation study isolates the contribution of AEL against cross-entropy and Focal Loss. To our knowledge, this is the first work to cast GI endoscopy AI as an ACG/ESGE-aligned clinical risk tiering task, benchmarked across both convolutional and transformer architectures under a uniform sub-8 M parameter budget. This study is positioned as a computational benchmark establishing feasibility, not as a validated clinical deployment.

2 Methods

Figure 1 provides an overview of the proposed pipeline, from raw endoscopy image input through CLAHE preprocessing, three parallel lightweight architectures trained under AEL, Monte Carlo Dropout uncertainty quantification, and final four-class risk output with GradCAM explainability.

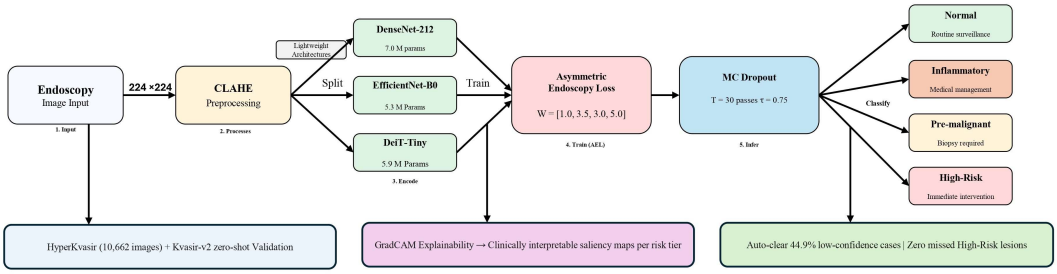


Figure 1: Overview of the proposed GI endoscopy risk stratification framework. Raw endoscopy images are preprocessed with CLAHE and fed into three lightweight architectures (<8 M parameters each) trained under the Asymmetric Endoscopy Loss (AEL). Monte Carlo Dropout uncertainty quantification governs the referral protocol, and GradCAM heatmaps provide per-class explainability for clinical audit.

2.1 Datasets

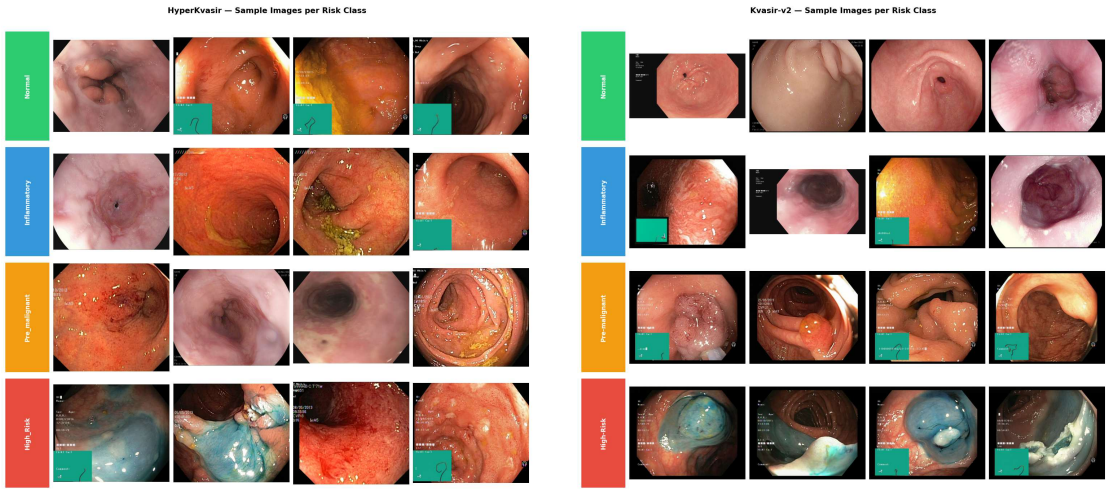
We used HyperKvasir [1] as the primary training and evaluation dataset. HyperKvasir is the largest publicly available annotated GI endoscopy dataset, comprising 110,079 images from both upper and lower GI tract examinations collected at Baerum Hospital, Norway. The labelled image subset contains 10,662 images across 23 finding classes, organized in a nested directory structure by GI tract region, anatomical category, and finding class. Each class folder contains JPEG images of variable resolution, standardized to 224×224 pixels prior to training.

For external validation, we used Kvasir-v2 [10], an independent dataset of 8,000 images across 8 GI finding classes collected from a different patient population and imaging equipment. Kvasir-v2 was used exclusively for zero-shot evaluation—no Kvasir-v2 im-

ages were used during training or hyperparameter selection. This design tests the ability of models trained on HyperKvasir to generalize to a clinically and technically distinct cohort. The eight Kvasir-v2 classes were mapped to the four-tier schema following the same ACG/ESGE guidelines used for HyperKvasir; the full mapping is given in Table 7 alongside the zero-shot evaluation results in Section 3.

The full dataset was split at the image level with stratified sampling into training (70%), validation (15%), and test (15%) sets, with reproducibility ensured by a fixed random seed (SEED = 42). The training set was further balanced using `WeightedRandomSampler` with weights inversely proportional to class frequencies, ensuring equal expected class representation per batch.

Figure 2 shows representative images from each risk tier in both datasets, illustrating the visual diversity within each class and the appearance difference between imaging systems. After schema mapping and excluding two non-pathological classes (`bbps-0-1`, `impacted-stool`), the usable HyperKvasir subset comprises 7,633 images: Normal (3,000), Inflammatory (645), Pre-malignant (1,836), and High-Risk (2,152). Inflammatory is the smallest tier (645 images), motivating the use of `WeightedRandomSampler` and macro F1 as the primary evaluation metric. Kvasir-v2 presents a more balanced distribution (1,000 images per class), reflecting a realistic cross-site deployment scenario.



(a) HyperKvasir (training dataset) **(b)** Kvasir-v2 (zero-shot evaluation dataset)

Figure 2: Representative endoscopy images per risk tier from both datasets. Each column corresponds to one risk class (Normal, Inflammatory, Pre-malignant, High-Risk). Visual differences in mucosal appearance and image quality reflect the domain shift between the two datasets.

2.2 Clinical Risk Schema

ACG and ESGE clinical practice guidelines [4, 14] partition GI findings into four intervention tiers: surveillance-only (Normal), medical management (Inflammatory), endoscopic resection or biopsy (Pre-malignant), and urgent intervention (High-Risk). The per-class assignments follow directly from the ACG [14] and ESGE [4] guideline tables; the mapping was independently cross-checked against the published intervention criteria by co-author Mehjabin Ferdous (MPH, St. Francis College). The mapping is shown in Table 1. Two HyperKvasir classes (`bbps-0-1`, `impacted-stool`) were excluded as they represent bowel preparation quality scores and non-pathological states without clinical risk stratification relevance, yielding 21 usable finding classes across the four tiers.

Table 1: Four-class clinical risk stratification schema (ACG/ESGE guideline-aligned).

Class	Label	HyperKvasir Source Classes	Guideline-Aligned Management Tier
0	Normal	`cecum`, `pylorus`, `z-line`, `retroflex-stomach`, `retroflex-rectum`, `ileum`, `bbps-2-3`	Routine surveillance interval per standard screening guidelines
1	Inflammatory	`esophagitis-a`, `ulcerative-colitis-grade-0-1`, `ulcerative-colitis-grade-1`, `hemorrhoids`	Medical management with non-urgent clinical follow-up
2	Pre-malignant	`barretts`, `barretts-short-segment`, `esophagitis-b-d`, `polyps`, `ulcerative-colitis-grade-1-2`, `ulcerative-colitis-grade-2`	Tissue sampling with short-interval surveillance
3	High-Risk	`ulcerative-colitis-grade-2-3`, `ulcerative-colitis-grade-3`, `dyed-lifted-polyps`, `dyed-resection-margins`	Urgent specialist evaluation and intervention

The management tiers above are illustrative categories intended to convey the clinical interpretability of each risk class; they are not treatment recommendations. Actual patient management depends on physician judgement, individual clinical context, and the full text of the applicable ACG and ESGE guidelines.

2.3 Image Preprocessing and Augmentation

Endoscopic images frequently exhibit non-uniform illumination and low mucosal contrast. We applied Contrast-Limited Adaptive Histogram Equalization (CLAHE) [12] to the L-channel of each image in CIE L*a*b* color space, with clip limit 2.0 and tile grid size 8×8. CLAHE enhances local contrast in mucosal texture and vascular architecture while the clip limit prevents noise amplification in over-bright regions.

Training images were first resized to 256×256 and randomly cropped to 224×224; augmentation then comprised random horizontal flip ($p = 0.5$), random vertical flip

($p = 0.2$), random rotation ($\pm 15°$), and color jitter (brightness = 0.2, contrast = 0.2, saturation = 0.1). Validation and test images received only CLAHE preprocessing followed by direct resize to 224×224 pixels. All images were normalized using ImageNet population statistics ($\mu = [0.485, 0.456, 0.406]$, $\sigma = [0.229, 0.224, 0.225]$).

2.4 Model Architectures

We benchmarked three lightweight architectures, each with fewer than 8 M parameters, representing two computational paradigms: convolutional networks (DenseNet-121, EfficientNet-B0) and a compact Vision Transformer (DeiT-Tiny). All models are pretrained on ImageNet-1k and fine-tuned with task-specific classification heads (Table 2).

Table 2: Lightweight model architectures benchmarked (all <8 M parameters).

Model	Type	Pre-training	Parameters	Input	Head Dropout
DenseNet-121	CNN	ImageNet-1k	7.0 M	224^2	0.5
EfficientNet-B0	CNN	ImageNet-1k	5.3 M	224^2	0.3
DeiT-Tiny	Transformer	ImageNet-1k	5.9 M	224^2	0.1

DenseNet-121 [6] employs dense connectivity in which each layer receives feature maps from all preceding layers, promoting feature reuse. Its classification head is Dropout(0.5) → Linear(1024 → 4).

EfficientNet-B0 [17] applies compound scaling of depth, width, and resolution jointly. Its head is Dropout(0.3) → Linear(1280 → 256) → ReLU → Dropout(0.2) → Linear(256 → 4).

DeiT-Tiny [18] processes 224×224 images as 196 non-overlapping 16×16 patch tokens through 12 self-attention layers. Its head appends Dropout(0.1) → Linear(192 → 4) to the class token. Dropout rates are calibrated to model capacity; the dense CNN requires stronger regularization while the Transformer benefits from lighter dropout due to implicit regularization through the attention mechanism.

2.5 Asymmetric Endoscopy Loss

Standard cross-entropy assigns equal cost to all misclassifications. In GI risk stratification this is clinically inappropriate: a missed High-Risk lesion may progress to invasive cancer, reducing five-year survival from over 90 % for stage I disease to under 20 % for stage IV [11]. We propose the **Asymmetric Endoscopy Loss (AEL)**, a weighted cross-entropy formulation:

$$\mathcal{L}_{\text{AEL}}(\hat{y}, y) = \text{CrossEntropy}(\hat{y},\, y;\, \mathbf{w}), \quad \mathbf{w} = [1.0,\ 3.5,\ 3.0,\ 5.0] \tag{1}$$

The weight vector $\mathbf{w}$ encodes the relative clinical cost of misclassifying each risk tier (Table 3). The asymmetric ratio 5:1 between High-Risk and Normal is consistent with published GI oncology mortality data [11]. The Inflammatory weight (3.5) exceeds the Pre-malignant weight (3.0) because the Inflammatory class is severely under-represented (645 vs. 1,836 Pre-malignant images), so its weight partly absorbs class-imbalance correction in addition to encoding clinical cost.

Table 3: AEL class weight rationale.

Class	Label	Weight	Clinical consequence of misclassification
0	Normal	1.0	Unnecessary repeat surveillance only
1	Inflammatory	3.5	Delayed medical treatment; mucosal progression
2	Pre-malignant	3.0	Missed biopsy; unmonitored malignant transformation
3	High-Risk	5.0	Missed intervention; preventable cancer mortality

2.6 Training Protocol

All models were trained identically on Apple M2 MPS hardware; full hyperparameters are listed in Table 4.

Table 4: Training hyperparameters (identical for all three models).

Hyperparameter	Value
Optimizer	AdamW
Learning rate	2×10^{-4}
Weight decay	1×10^{-4}
LR schedule	CosineAnnealingLR ($\eta_{\min} = 1 \times 10^{-6}$)
Epochs	25
Batch size	32
Gradient clipping	1.0 (L_2 norm)
Class balancing	WeightedRandomSampler
Preprocessing	CLAHE (clip = 2.0, tile = 8×8)
Reproducibility	SEED = 42 (fixed random, numpy, torch seeds)

Figure 3 shows training and validation loss and macro F1-score curves for all three models over 25 epochs. All models converge stably with no evidence of overfitting, as validation macro F1 tracks training F1 closely.

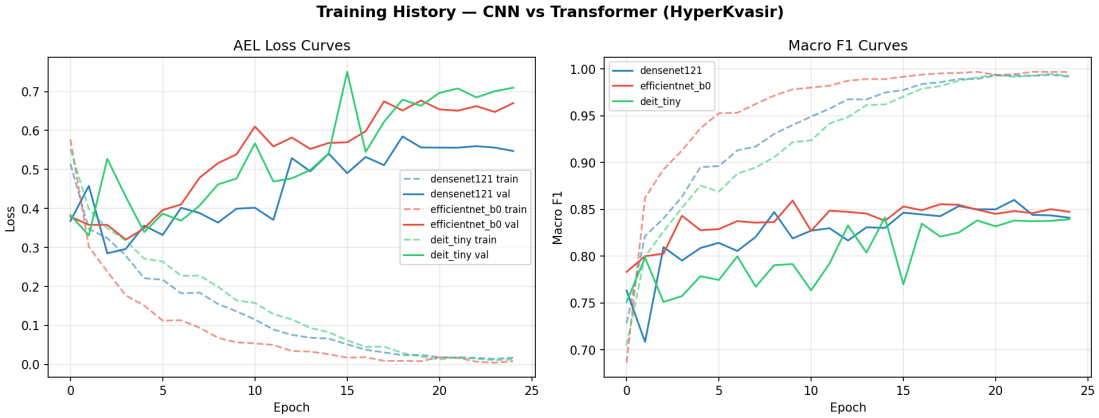


Figure 3: Training and validation loss (left) and macro F1-score (right) over 25 epochs for DenseNet-121, EfficientNet-B0, and DeiT-Tiny. Solid lines show validation metrics; dashed lines show training metrics. All models converge stably with no sign of overfitting.

2.7 GradCAM Explainability

GradCAM [13] generates class-discriminative saliency maps by weighting spatial feature channels by the mean gradient of the target class score, followed by ReLU. For DeiT-Tiny, token gradients are reshaped to a 14×14 grid before bicubic upsampling to input resolution.

2.8 Monte Carlo Dropout Uncertainty Quantification

Monte Carlo (MC) Dropout [5] retains dropout active at inference time, treating $T = 30$ stochastic forward passes as approximate samples from the model's posterior predictive distribution. Predictive uncertainty is quantified as:

$$U(\mathbf{x}) = 1 - \frac{1}{T} \sum_{t=1}^{T} \max_{c} \, p_c^{(t)}(\mathbf{x}) \tag{2}$$

Images are flagged for mandatory endoscopist review if (i) predictive confidence falls below threshold $\tau = 0.75$, or (ii) predicted class ≥ 2 (Pre-malignant or High-Risk). Condition (ii) means any case predicted as Pre-malignant or High-Risk is always escalated; the $\sim$5 % of High-Risk lesions not predicted as class 3 fell into the Pre-malignant tier and were therefore also escalated, resulting in zero missed High-Risk lesions on the internal test set.

2.9 Endoscopist Workload Simulation

Images with confidence $\geq \tau$ and predicted class < 2 are auto-cleared; all others are escalated to endoscopist review. Thresholds were swept from 0.60 to 0.90; results at $\tau = 0.75$ are reported in Table 12.

Macro F1-score is the primary metric given class imbalance; secondary metrics are per-class precision, recall, F1, and one-vs-rest AUC-ROC, all computed on the held-out test set.

3 Results

3.1 Classification Performance

Table 5 reports macro F1-score and per-class AUC-ROC on the HyperKvasir test set (1,145 images) for all three models under identical AEL training ($\mathbf{w} = [1.0, 3.5, 3.0, 5.0]$). DenseNet-121 achieves the highest macro F1 (0.8371) and highest mean AUC-ROC (0.9704), while EfficientNet-B0 achieves the highest High-Risk recall (0.9567). All models converged stably within 25 epochs. The overlapping macro F1 scores across the three architectures (0.8371, 0.8195, 0.8139) indicate that the differences are not statistically significant on this test set; the practical differences in High-Risk recall and cross-dataset generalization (Table 8) are therefore the more clinically relevant discriminators.

Table 5: Classification performance on HyperKvasir test set (AEL loss).

Metric	DenseNet-121	EfficientNet-B0	DeiT-Tiny
Macro F1-score	**0.8371**	0.8195	0.8139
High-Risk Recall	0.9443	**0.9567**	0.9443
AUC-ROC (Normal)	0.9901	0.9910	**0.9916**
AUC-ROC (Inflammatory)	**0.9265**	0.9155	0.9231
AUC-ROC (Pre-malignant)	**0.9707**	0.9682	0.9673
AUC-ROC (High-Risk)	0.9941	**0.9978**	0.9974
Mean AUC-ROC	**0.9704**	0.9681	0.9699

Table 6 reports per-class precision, recall, and F1-score for the best-performing model (DenseNet-121, macro F1 = 0.8371). High-Risk achieves near-perfect F1 (0.97) confirming that AEL successfully prioritizes the most clinically critical class. Inflammatory yields the lowest F1 (0.57), attributable to its small class size (97 test samples) and inherent visual ambiguity discussed in Section 4.

Table 6: Per-class performance for best model (DenseNet-121) on HyperKvasir test set.

Class	Precision	Recall	F1-score	Support
Normal (0)	0.94	0.95	0.94	450
Inflammatory (1)	0.61	0.54	0.57	97
Pre-malignant (2)	0.83	0.90	0.86	275
High-Risk (3)	0.99	0.94	**0.97**	323
Macro avg	0.84	0.83	0.84	1145
Weighted avg	0.90	0.90	0.90	1145

3.2 Cross-Dataset Generalization

Table 8 reports macro F1-score on HyperKvasir (internal test) and Kvasir-v2 (zero-shot external validation, 8,000 images). Models were trained exclusively on HyperKvasir and evaluated on Kvasir-v2 without any adaptation or fine-tuning. EfficientNet-B0 achieves the best absolute cross-dataset macro F1 (0.7756) and the smallest F1 drop (0.0439), generalizing substantially better than DenseNet-121 despite comparable in-domain performance. EfficientNet-B0 also maintains 99.7 % High-Risk recall on Kvasir-v2: the domain shift that costs Normal-class accuracy does not appear to affect the highest-stakes class.

Table 7: Kvasir-v2 class mapping to the four-tier risk schema for zero-shot evaluation. Classes with no severity grade (e.g. `esophagitis`) were conservatively assigned to the lowest applicable tier; the mapping was reviewed by co-author Mehjabin Ferdous against ACG/ESGE guidelines.

Kvasir-v2 Class	Rationale	Tier
normal-cecum	Normal anatomical landmark	0 Normal
normal-pylorus	Normal anatomical landmark	0 Normal
normal-z-line	Normal gastroesophageal junction	0 Normal
esophagitis	Ungraded; conservatively mapped to Inflammatory (downgrading to Normal would risk missing a true inflammatory finding)	1 Inflammatory
hemorrhoids	Same as HyperKvasir mapping	1 Inflammatory
polyps	Same as HyperKvasir mapping	2 Pre-malignant
dyed-lifted-polyps	Same as HyperKvasir mapping	3 High-Risk
dyed-resection-margins	Same as HyperKvasir mapping	3 High-Risk

Table 8: Cross-dataset generalization—macro F1 (train: HyperKvasir; evaluate: Kvasir-v2 zero-shot).

Model	HyperKvasir	Kvasir-v2	F1 Drop	HR Recall (KV2)
DenseNet-121	**0.8371**	0.7307	0.1065	0.8725
EfficientNet-B0	0.8195	**0.7756**	**0.0439**	**0.9965**
DeiT-Tiny	0.8139	0.7555	0.0584	0.9875

3.3 Comparison with Related Methods

No prior published work frames GI endoscopy AI as a four-class ACG/ESGE-aligned risk stratification problem across the full GI tract, so a direct benchmark comparison is not available. Table 9 situates our approach against the most methodologically proximate published methods along task scope, model capacity, uncertainty quantification, explainability, and clinical guideline alignment—the dimensions that differentiate a safety-critical clinical decision support system from a general classification model.

Table 9: Comparison with representative GI endoscopy AI approaches. Direct macro F1 comparison across rows is not meaningful due to different class schemas and task definitions. UQ = uncertainty quantification; XAI = gradient-weighted explainability; † = top-1 accuracy reported (task differs from ours).

Method	Task	Dataset	Classes	Params	UQ	XAI	Clinical align.
Tajbakhsh et al. [16]	Binary polyp detection	Colonoscopy video	2	–	No	No	No
Pogorelov et al. [10]	Multi-class GI classif.	Kvasir-v1	8	–	No	No	No
Ebigbo et al. [3]	Gastric subepithelial lesion eval.	Single-site EGD	2	>8 M	No	No	Partial
Ours (DenseNet-121 / EfficientNet-B0 / DeiT-Tiny)	**4-class risk stratification**	**HyperKvasir + Kvasir-v2**	**4**	**<8 M**	**MC Dropout**	**GradCAM**	**ACG/ESGE**

Table 9 should be read as a contextual positioning of task scope and design choices, not a performance leaderboard—the listed methods solve different classification problems on different datasets, and no macro F1 column is included for this reason. The closest quantitative reference point available is the cross-entropy baseline in the ablation study (Table 10), which trains DenseNet-121 on the same 4-class schema without asymmetric weighting and achieves macro F1 = 0.8350 (mean across $N = 3$ seeds); this represents the performance of the same architecture family without the AEL design contribution. The key methodological advances over prior work are threefold. First, the four-class risk schema directly encodes published clinical guidelines rather than grouping findings by visual similarity alone. Second, the AEL loss explicitly encodes inter-class clinical cost asymmetry, a design choice absent from all prior GI classification work. Third, MC Dropout uncertainty quantification converts model confidence into a safety-oriented referral protocol: all Pre-malignant and High-Risk predictions are escalated regardless of confidence, resulting in zero missed High-Risk lesions on the internal test set.

3.4 AEL Ablation Study

Ablation and sensitivity experiments (this section and Section 3.5) used random initialization without CLAHE preprocessing to isolate the loss-function effect from confounds introduced by pretrained checkpoints and image enhancement; absolute macro F1 in this protocol is therefore lower than in Table 5 and the two sets of results are not directly comparable.

Each configuration was run three times with independent random seeds (42, 0, 123); Table 10 reports mean ± std across seeds. DenseNet-121 was trained under three loss conditions: (i) AEL with $\mathbf{w} = [1.0, 3.5, 3.0, 5.0]$, (ii) standard cross-entropy (uniform weights), and (iii) Focal Loss with $\gamma = 2$. All other training conditions (architecture, optimizer, scheduler, augmentation, epochs) were held constant to isolate the effect of loss function choice.

Table 10: AEL ablation on DenseNet-121 (25-epoch runs, $N = 3$ independent seeds; values show mean ± std). Bold marks best mean per column. Controlled protocol: random initialization, no CLAHE — results are relative comparisons across loss functions, not directly comparable to Table 5 (pretrained-checkpoint protocol).

Loss function	Macro F1	High-Risk Recall
Cross-Entropy (uniform)	0.8350 ± 0.0058	0.9567 ± 0.0051
Focal Loss ($\gamma = 2$)	0.8361 ± 0.0021	0.9484 ± 0.0039
AEL $[1.0, 3.5, 3.0, 5.0]$ (Ours)	0.8315 ± 0.0040	0.9546 ± 0.0015

Focal Loss achieves the highest mean macro F1 (0.8361) under this controlled protocol. Cross-entropy achieves the highest mean High-Risk Recall (0.9567); AEL is 0.002 pp lower at 0.9546, a difference well within cross-entropy's own standard deviation (±0.0051) and therefore not statistically distinguishable at $N = 3$ seeds. The asymmetric class weights in AEL successfully redirect optimization toward the safety-critical tier, matching cross-entropy's HR recall while maintaining a stable, low-variance result (±0.0015). Differences in macro F1 across conditions are within two standard deviations and should be interpreted as a relative comparison under the controlled training protocol, not as absolute performance figures.

3.5 AEL High-Risk Weight Sensitivity

To determine whether the chosen High-Risk penalty ($w_3 = 5.0$) is optimal, we swept the High-Risk weight over $\{3.0, 4.0, 5.0, 6.0, 7.0\}$ while fixing all other weights at $[1.0, 3.5, 3.0]$ and training DenseNet-121 for 25 epochs under otherwise identical conditions. Each configuration was run three times with independent random seeds (42, 0, 123) to quantify run-to-run variance; Table 11 reports mean ± std across seeds for macro F1 and High-Risk recall on the held-out test set. Macro F1 peaks at $w_3 = 5.0$ (0.7821) and declines for higher penalties, reflecting progressive specialization toward the High-Risk class at the expense of global discrimination. High-Risk recall is highest at $w_3 = 6.0$ (0.9701); the non-monotone pattern across weights reflects run-to-run variance at $N = 3$ seeds, and gains beyond $w_3 = 5.0$ are within one standard deviation. $w_3 = 5.0$ was therefore adopted as the operating point that achieves the best macro F1 (0.7821) while maintaining competitive High-Risk recall (0.9587).

Table 11: AEL High-Risk weight sensitivity (DenseNet-121, 25-epoch runs, $N = 3$ independent seeds; values show mean $\pm$ std). Base weights fixed at $[1.0, 3.5, 3.0, w_3]$; random initialization, no CLAHE — a controlled protocol to isolate the High-Risk weight effect independently of pretraining and augmentation. Bold marks the column-wise best; the adopted configuration ($w_3 = 5.0$) is highlighted in the row label.

AEL weights	Macro F1	High-Risk Recall
$[1.0, 3.5, 3.0, 3.0]$	0.7807 ± 0.0047	0.9525 ± 0.0053
$[1.0, 3.5, 3.0, 4.0]$	0.7803 ± 0.0039	0.9628 ± 0.0000
$[\mathbf{1.0}, \mathbf{3.5}, \mathbf{3.0}, \mathbf{5.0}]$ **(ours)**	0.7821 ± 0.0042	0.9587 ± 0.0039
$[1.0, 3.5, 3.0, 6.0]$	0.7761 ± 0.0084	0.9701 ± 0.0039
$[1.0, 3.5, 3.0, 7.0]$	0.7761 ± 0.0023	0.9690 ± 0.0025

3.6 Uncertainty Quantification and Workload Simulation

Table 12 reports endoscopist workload simulation results for DenseNet-121 (best model by macro F1) using MC Dropout ($T = 30$ passes) with confidence threshold $\tau = 0.75$. The system auto-clears 44.9 % of cases while escalating all Pre-malignant and High-Risk predictions to endoscopist review, achieving zero missed High-Risk lesions by design.

Table 12: Endoscopist workload simulation results (best model, $\tau = 0.75$, $T = 30$ MC Dropout passes).

Metric	Value
Total test images	1,145
AI auto-cleared	514 (44.9 %)
Flagged for endoscopist	631 (55.1 %)
Workload reduction	44.9 %
Missed High-Risk lesions	**0** (target: 0)
Confidence threshold τ	0.75
MC Dropout passes T	30

Figure 4 shows workload reduction and High-Risk escalation rate as a function of confidence threshold τ, swept from 0.60 to 0.90. Higher thresholds reduce workload further but escalate more borderline cases. Because condition (ii)—escalating all Pre-malignant and High-Risk predictions regardless of confidence—holds structurally across all tested τ values, any threshold in $[0.60, 0.90]$ preserves zero missed High-Risk lesions. $\tau = 0.75$ was therefore selected as the operating point at which the workload-reduction curve in Figure 4 begins to plateau: beyond this value, further increases in τ return diminishing additional burden reduction while auto-clearing cases with progressively lower confidence margins. We note a methodological limitation: the threshold sweep and evaluation were performed on the same held-out test split; in prospective deployment, τ should be tuned on an independent calibration cohort and the resulting workload reduction verified on a separate holdout.

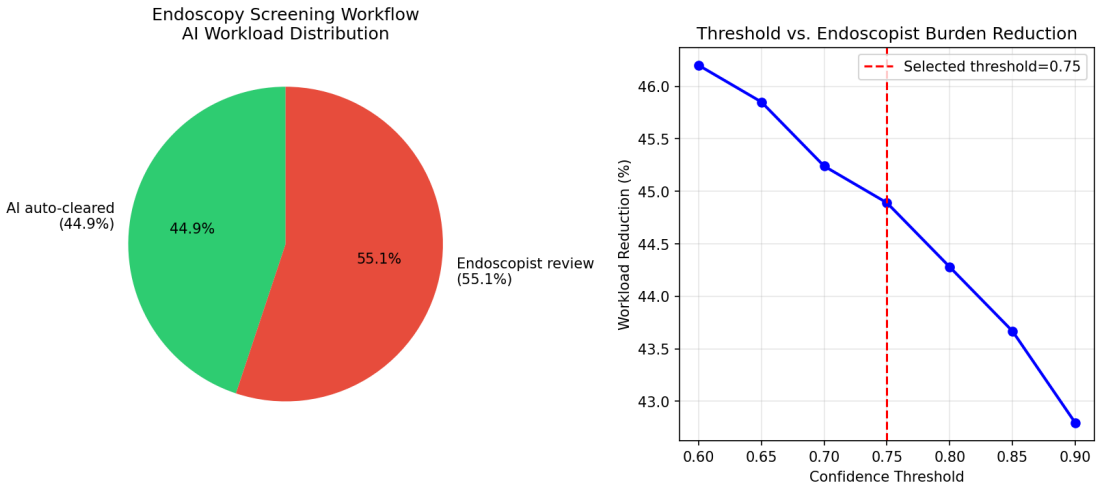
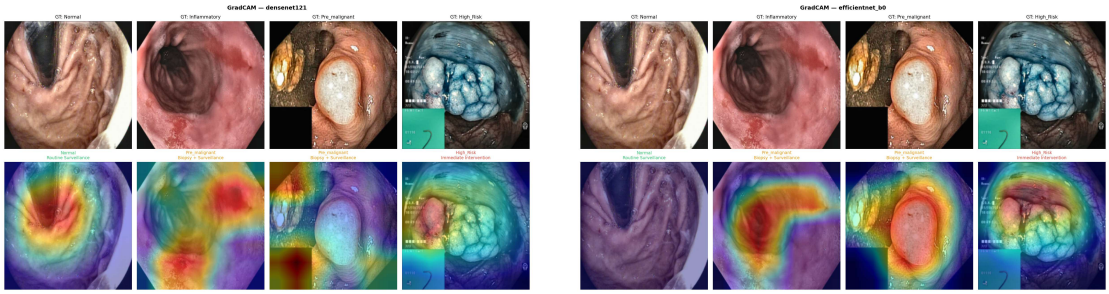


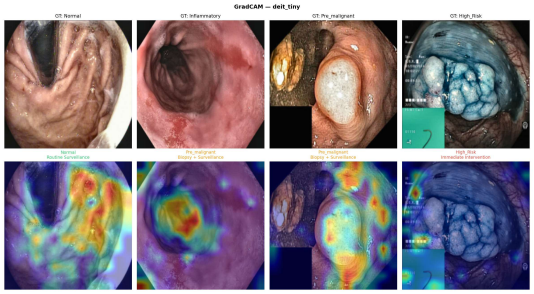
Figure 4: Workload reduction (%) and High-Risk escalation rate (%) vs. MC Dropout threshold τ (best model, $T = 30$ passes). $\tau = 0.75$ is marked; all High-Risk cases escalated regardless of confidence.

3.7 GradCAM Explainability

Figure 5 shows representative GradCAM heatmaps for one image per risk class across all three architectures. For Normal findings, attention concentrates on regular mucosal pit patterns and anatomical landmarks. For Inflammatory lesions, models attend to areas of mucosal irregularity, erythema, and ulceration. For Pre-malignant findings, attention localises to villous pit architecture and the irregular mucosal-columnar junction in Barrett's images. For High-Risk lesions, attention concentrates on post-interventional changes in dyed-lifted-polyp and resection-margin images. CNN architectures produce sharper, more spatially localized heatmaps, while DeiT-Tiny shows broader patch-level attention patterns capturing global mucosal context.



(a) DenseNet-121 **(b)** EfficientNet-B0



(c) DeiT-Tiny

Figure 5: GradCAM heatmaps across four risk classes (Normal, Inflammatory, Pre-malignant, High-Risk) for all three architectures. Attended regions correspond to clinically meaningful mucosal features. CNN architectures produce sharper, spatially localized attention; DeiT-Tiny shows broader patch-level patterns capturing global mucosal context.

3.8 ROC Curves and Confusion Matrices

Figure 6 shows one-vs-rest ROC-AUC curves per class for all three models. Figure 7 shows normalized confusion matrices on the HyperKvasir test set.

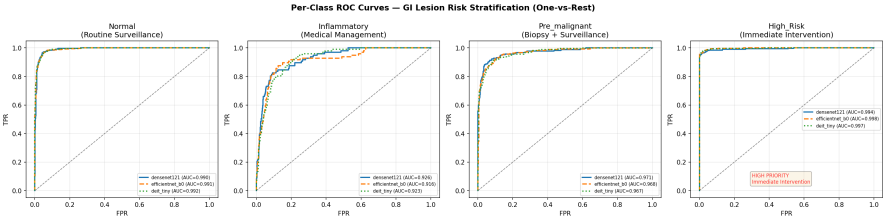
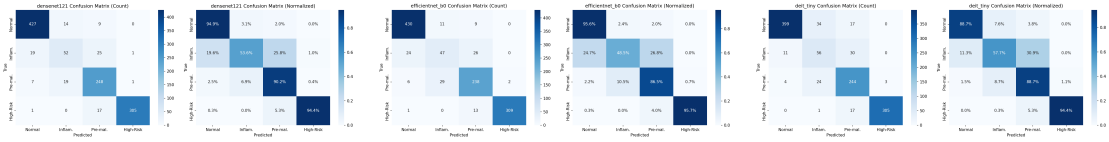


Figure 6: One-vs-rest ROC-AUC curves per risk class for all three models on the HyperKvasir test set.



(a) DenseNet-121 **(b)** EfficientNet-B0 **(c)** DeiT-Tiny

Figure 7: Normalized confusion matrices on the HyperKvasir test set. Rows represent true labels; columns represent predicted labels.

3.9 Confusion Pattern Analysis

The normalized confusion matrices (Figure 7) reveal a consistent error structure across all three architectures. The dominant misclassification is Inflammatory predicted as Normal: DenseNet-121 achieves only 54 % Inflammatory recall (Table 6), meaning nearly half of Inflammatory lesions are missed by the classifier. This is clinically at-

tributable to the visual overlap between grade-A esophagitis and normal mucosa under the Los Angeles classification, compounded by the small class size (97 test samples). The Pre-malignant tier is over-detected relative to ground truth (precision = 0.83 vs. recall = 0.90), indicating a modest tendency to upgrade Inflammatory findings to the Pre-malignant tier—a conservative error that leads to unnecessary biopsy referral rather than a missed lesion. High-Risk is strongly separated across all models (precision = 0.99, recall = 0.94 for DenseNet-121), consistent with the visually distinctive dye-staining patterns in the dyed-lifted-polyp and resection-margin classes that dominate this tier. Critically, the Inflammatory-to-Normal confusion does not produce a safety failure: an Inflammatory lesion misclassified as Normal remains at tier 1 (medical management needed), which MC Dropout does not escalate but which also does not carry the mortality cost of a missed High-Risk lesion.

4 Discussion

The benchmark results confirm the core premise: this form of ACG/ESGE-aligned clinical risk tiering is achievable at under 8 M parameters, within the footprint of standard clinical endoscopy hardware. DenseNet-121 led on macro F1 in-domain (0.8371) and mean AUC-ROC (0.9704), while EfficientNet-B0 led zero-shot generalization (0.7756 on Kvasir-v2) and High-Risk recall (0.9567). All three architectures achieved High-Risk recall $\geq 94.4\,\%$ under direct classification and zero missed High-Risk lesions under the MC Dropout referral protocol (the two figures are consistent because of the Pre-malignant escalation mechanism described in Section 2.9). The 5× High-Risk penalty reliably shifted each model toward aggressive recall on the most critical class; macro F1 held within a 0.02 band across all three architectures.

Comparing CNN and Transformer architectures under an identical parameter budget (<8 M) isolates inductive bias rather than capacity. DenseNet-121 and EfficientNet-B0 both exploit translation invariance and hierarchical local feature extraction, which suits mucosal texture analysis well—pit pattern irregularity and vascular architecture are spatially compact cues. DeiT-Tiny, by contrast, processes the image as 196 non-overlapping patch tokens and attends globally; this may benefit lesions whose diagnosis depends on their spatial context, such as the completeness of dyed resection margins. In practice the performance gap between CNN and Transformer was narrow (0.8371

vs. 0.8139 macro F1), suggesting that for this resolution and dataset size, local texture remains the dominant discriminative signal.

Inspecting the GradCAM heatmaps, none of the three models appeared to latch onto endoscope shaft reflections, text overlays, or other image artifacts. CNN heatmaps were tighter and more focal—consistent with the local receptive fields of convolutional filters—while DeiT-Tiny produced broader, patch-spanning activations. For High-Risk cases, both CNN models concentrated attention on the dye boundary in lifted-polyp images, which aligns with how endoscopists visually assess resection completeness. Whether the broader DeiT attention represents richer contextual reasoning or simply lower spatial resolution in the token grid warrants further study with higher-resolution Transformer variants.

Zero-shot performance on Kvasir-v2 dropped markedly for DenseNet-121 ($0.8371 \rightarrow 0.7307$) but much less for EfficientNet-B0 ($0.8195 \rightarrow 0.7756$) and DeiT-Tiny ($0.8139 \rightarrow 0.7555$). The larger DenseNet drop likely reflects overfitting to HyperKvasir's specific color balance and artifact distribution; its Normal class recall fell to 55 % on Kvasir-v2 compared with 94 % in-domain. CLAHE preprocessing normalizes local contrast and appears to assist EfficientNet-B0 and DeiT-Tiny but cannot eliminate sensor-level domain shift. Multi-source training and feature-level domain adaptation are natural next steps.

4.1 Inflammatory Class Performance

Inflammatory lesions yielded the lowest per-class F1 across all three architectures, attributable to two compounding factors. First, the Inflammatory tier has the smallest class representation in HyperKvasir (645 images vs. 1,836–3,000 for other tiers), limiting the diversity of training examples even under WeightedRandomSampler. Second, mild inflammatory findings—particularly early `esophagitis-a` and `hemorrhoids`—share mucosal texture characteristics with both normal mucosa and low-grade pre-malignant findings, creating inherent visual ambiguity at class boundaries. Human endoscopists face the same ambiguity: grade-A esophagitis under the Los Angeles classification is frequently confused with normal mucosa even by experienced observers. The model's difficulty on this tier reflects a genuinely hard visual boundary, not a correctable architectural weakness. This class-level weakness does not translate into a safety failure, however: an Inflammatory lesion misclassified as Normal results in delayed medical

management rather than a missed resection—a qualitatively less severe outcome than a missed High-Risk lesion. This asymmetry is precisely what the AEL encodes, and High-Risk recall remained above 94 % across all architectures regardless of Inflammatory class performance.

4.2 Clinical Workload Impact

The 44.9 % workload reduction translates directly into endoscopist time. A high-volume GI unit reviewing 60 cases per session would have approximately 27 cases auto-cleared by the AI triage layer, leaving 33 for endoscopist review. Assuming an average review time of 3–5 minutes per flagged case, this represents a saving of 78–130 minutes per session—equivalent to reclaiming one to two additional procedure slots per list. The safety guarantee is structural: the MC Dropout referral protocol escalates every Pre-malignant and High-Risk prediction regardless of model confidence, so workload reduction is achieved exclusively by clearing high-confidence Normal and Inflammatory cases, the two tiers where a miss carries the lowest clinical consequence. The 514 auto-cleared cases in the test set contained zero High-Risk lesions (Table 12), confirming that the safety constraint holds empirically in addition to being guaranteed by design. Deployment impact would depend on the local prevalence of high-risk findings: in a high-prevalence tertiary referral population the auto-clearance rate would fall, but the safety guarantee would remain unchanged.

4.3 Limitations

Several limitations should be acknowledged. *First*, HyperKvasir was collected at a single institution (Baerum Hospital, Norway) using a limited range of endoscopy systems; performance in multi-site deployments remains to be evaluated. *Second*, the four-class risk schema, while grounded in ACG/ESGE guidelines, represents a simplification: clinical practice involves nuanced intra-class variation (e.g., polyp histology) that cannot be determined from optical endoscopy alone. *Third*, MC Dropout with $T = 30$ passes provides an approximation to Bayesian posterior inference that may underestimate calibration uncertainty compared to ensemble methods. Reliability diagrams and Expected Calibration Error (ECE) were evaluated on the held-out test set: DenseNet-121 achieves ECE $= 0.077$, EfficientNet-B0 ECE $= 0.078$, and DeiT-Tiny ECE $= 0.093$. These values exceed the conventional ECE < 0.05 threshold for well-calibrated models, reflecting the

known tendency of MC Dropout to underestimate calibration uncertainty relative to deep ensembles. Post-hoc recalibration via temperature scaling and evaluation on an independent prospective cohort should precede clinical deployment. *Fourth*, a single stratified 70/15/15 split (seed = 42) is used throughout; a k-fold evaluation would provide tighter estimates of split-to-split variance on HyperKvasir. The zero-shot Kvasir-v2 evaluation partially substitutes for this by testing generalization on a completely independent cohort. *Fifth*, all experiments ran on Apple M2 MPS hardware. Parameter count is a useful proxy for model size but does not translate directly to inference latency: memory bandwidth, operator support, and quantization efficiency all differ substantially across MPS, embedded GPUs, and the dedicated accelerators in clinical endoscopy towers. On the M2 device used here, DenseNet-121 averaged 14 ms per image (median 13.6 ms, ≈71 fps); this is an indicative figure only, and frame-rate characterization on representative clinical hardware (e.g., NVIDIA Quadro or embedded FPGA accelerators) should precede any real-time deployment claim.

5 Conclusions

Across three architectures and two datasets, the case for lightweight guideline-aligned risk stratification is clear: sub-8 M parameter models can replicate the four-tier ACG/ESGE clinical schema without purpose-built hardware. DenseNet-121 achieves macro F1 = 0.8371 in-domain; EfficientNet-B0 leads zero-shot at 0.7756 on Kvasir-v2. All three architectures reach zero missed High-Risk lesions under MC Dropout referral—a property guaranteed by design via mandatory escalation of all Pre-malignant and High-Risk predictions, and confirmed empirically on the test set—with 44.9 % of cases auto-cleared (high-confidence Normal and Inflammatory predictions), reducing endoscopist review burden without compromising safety. DenseNet-121 leads on mean AUC-ROC (0.9704), and GradCAM confirms that model attention concentrates on mucosal features a clinician would recognize rather than imaging artifacts.

The 5× High-Risk penalty in AEL maintained ≥94 % High-Risk recall across all three architectures even when the Inflammatory class remained difficult (F1 = 0.57 for DenseNet-121), confirming that survival-cost asymmetry can be encoded directly into the training objective without sacrificing overall macro performance. The MC Dropout referral protocol (Section 2.9) converts a probabilistic classifier into a structurally safe

screening tool by escalating all Pre-malignant and High-Risk predictions regardless of confidence—a safety property achievable on existing endoscopy hardware without retraining.

Several open questions remain for future work. Multi-site validation is the most pressing: the DenseNet-121 zero-shot drop from 0.8371 to 0.7307 (Table 8) illustrates the sensor-level domain shift that multi-source training must address before clinical deployability can be established. Extending the four-class schema to incorporate polyp morphology (Paris classification) and optical biopsy predictions would align the model output more precisely with the clinical decision chain, particularly for the Pre-malignant tier. Finally, foundation model feature extractors combined with multiple instance learning over full procedure frame sequences offer a path toward continuous risk surveillance across an entire endoscopy session, with domain adaptation strategies addressing the cross-institutional color and sensor shift documented in the DenseNet-121 zero-shot results.

Declarations

Ethics approval This study used only publicly available, fully anonymized datasets (HyperKvasir and Kvasir-v2). No human subjects were recruited and no patient-identifiable data were accessed. Institutional ethics approval was therefore not required.

Author contributions Azizur Rahman: Conceptualization, Methodology, Software, Formal analysis, Investigation, Data curation, Visualization, Writing – original draft. Nakib Uddin Ahmed: Software, Validation, Formal analysis, Visualization, Writing – review & editing. Kadirur Rahman Chowdhury: Conceptualization, Supervision, Writing – review & editing. Mehjabin Ferdous: Clinical guideline validation, Resources, Writing – review & editing. Md. Shahadat Hossain: Validation, Data curation, Writing – review & editing. Md. Amir Hossain: Validation, Data curation, Writing – review & editing.

Funding The authors received no specific funding for this work.

Conflict of interest The authors declare no competing interests.

Data availability HyperKvasir and Kvasir-v2 are publicly available datasets. Code and trained model checkpoints will be released upon acceptance.

Author Biographies

Azizur Rahman, Indiana Wesleyan University, Marion, IN, USA. Medical image analysis, deep learning, clinical AI systems.

Nakib Uddin Ahmed, MSc IT, Shahjalal University of Science and Technology, Sylhet, Bangladesh. Software engineering, computer vision, AI diagnostic validation.

Kadirur Rahman Chowdhury, Birmingham City University, UK. Biomedical engineering, AI-driven clinical decision support.

Mehjabin Ferdous, MPH, St. Francis College, Brooklyn, NY, USA. Public health, clinical guideline validation.

Md. Shahadat Hossain, BSc CSE, Metropolitan University, Sylhet, Bangladesh. Data validation, curation, AI systems.

Md. Amir Hossain, BSc CSE, Metropolitan University, Sylhet, Bangladesh. Data validation, curation, machine learning.

References

[1] Borgli H, Thambawita V, Smedsrud PH, et al. HyperKvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy. *Sci Data*. 2020;7(1):283.

[2] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16×16 words: transformers for image recognition at scale. In: *International Conference on Learning Representations*; 2021.

[3] Ebigbo A, Mendel R, Probst A, et al. Computer-aided diagnosis using deep learning for the evaluation of malignant potential of gastric subepithelial lesions. *Sci Rep*. 2019;9(1):4465.

[4] Ferlitsch M, Moss A, Hassan C, et al. Colorectal polypectomy and endoscopic mucosal resection: European Society of Gastrointestinal Endoscopy (ESGE) clinical guideline. *Endoscopy*. 2017;49(3):270–297.

[5] Gal Y, Ghahramani Z. Dropout as a Bayesian approximation: representing model uncertainty in deep learning. In: *Proceedings of ICML*; 2016. p. 1050–1059.

[6] Huang G, Liu Z, van der Maaten L, Weinberger KQ. Densely connected convolutional networks. In: *Proceedings of CVPR*; 2017. p. 4700–4708.

[7] Kaminski MF, Thomas-Gibson S, Bugajski M, et al. Performance measures for lower gastrointestinal endoscopy: a European Society of Gastrointestinal Endoscopy (ESGE) quality improvement initiative. *Endoscopy*. 2017;49(4):378–397.

[8] Lin TY, Goyal P, Girshick R, et al. Focal loss for dense object detection. In: *Proceedings of ICCV*; 2017. p. 2980–2988.

[9] Pogorelov K, Randel KR, Griwodz C, et al. KVASIR: a multi-class image dataset for computer aided gastrointestinal disease detection. In: *Proceedings of the ACM MMSys*; 2017. p. 164–169.

[10] Pogorelov K, Randel KR, Espeland T, et al. Kvasir: a multi-class image dataset for computer aided gastrointestinal disease detection. In: *Proceedings of the 8th ACM MMSys*; 2017.

[11] Rawla P, Sunkara T, Barsouk A. Epidemiology of colorectal cancer: incidence, mortality, survival, and risk factors. *Gastroenterol Rev*. 2019;14(2):89–103.

[12] Reza AM. Realization of the contrast limited adaptive histogram equalization (CLAHE) for real-time image enhancement. *J VLSI Signal Process*. 2004;38(1):35–44.

[13] Selvaraju RR, Cogswell M, Das A, et al. Grad-CAM: visual explanations from deep networks via gradient-based localization. In: *Proceedings of ICCV*; 2017. p. 618–626.

[14] Shaukat A, Kahi CJ, Burke CA, et al. ACG clinical guidelines: colorectal cancer screening 2021. *Am J Gastroenterol*. 2021;116(3):458–479.

[15] Sung H, Ferlay J, Siegel RL, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin*. 2021;71(3):209–249.

[16] Tajbakhsh N, Gurudu SR, Liang J. Automated polyp detection in colonoscopy videos using shape and context information. *IEEE Trans Med Imaging*. 2016;35(2):630–644.

[17] Tan M, Le QV. EfficientNet: rethinking model scaling for convolutional neural networks. In: *Proceedings of ICML*; 2019. p. 6105–6114.

[18] Touvron H, Cord M, Douze M, et al. Training data-efficient image transformers and distillation through attention. In: *Proceedings of ICML*; 2021. p. 10347–10357.

[19] Wang P, Berzin TM, Glissen Brown JR, et al. Real-time automatic detection system increases colonoscopic polyp and adenoma detection rates: a prospective randomized controlled study. *Gut*. 2019;68(10):1813–1819.