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.

CBAM-Enhanced MaxViT-Tiny for Gastrointestinal Endoscopy Image Classification: Robustness, Explainability, and External Validation

Sadek Al Prince

sadekprince09@gmail.com

International Islamic University Chittagong

Israt Fatema Shorna

International Islamic University Chittagong

Md Mehedi Hasan

Independent University Bangladesh

Mohammad Ahanaf Akif

Daffodil International University

Aziz Misher Mishu

International Islamic University Chittagong

Gazi Mehraj Sakib

International Islamic University Chittagong

Article

Keywords:

Posted Date: November 9th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11408118/v1

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

Additional Declarations: No competing interests reported.

CBAM-Enhanced MaxViT-Tiny for Gastrointestinal Endoscopy Image Classification: Robustness, Explainability, and External Validation

Sadek Al Prince[1*], Israt Fatema Shorna[1†], Md Mehedi Hasan[2], Mohammad Ahanaf Akif[3], Aziz Misher Mishu[1], Gazi Mehraj Sakib[1]

[1*]Department of Computer Science and Engineering, International Islamic University Chittagong, Bangladesh.
[2]Department of Computer Science and Engineering, Independent University Bangladesh, Bangladesh.
[3]Department of Software Engineering (Major in Cyber Security), Daffodil International University, Bangladesh.

*Corresponding author(s). E-mail(s): sadekprince09@gmail.com;
Contributing authors: isratfatema2356@gmail.com; mehedi.mh6522@gmail.com; mohammed15-4256@diu.edu.bd; azizmishermishu952@gmail.com; gazimehrajsakib@gmail.com;
†These authors contributed equally to this work.

Abstract

Automated classification of the huge number of images generated during gastrointestinal (GI) endoscopy is required for reliable and interpretable diagnosis but there is concern about the vulnerability of deep models to adversarial perturbations. In the current study, a MaxViT-Tiny backbone and a Convolutional Block Attention Module (CBAM) head are used for 8-class classification of the Kvasir endoscopic images. After deleting the duplicate images, there are 3,996 images, which were split into 3 sets: 70% as the training set, 15% as the validation set, and 15% as the test set; the model was compared with 5 backbones with the same CBAM head under 3 random seeds. This resulted in the highest test accuracy of 94.33%, a macro F1 score of 0.9431, and a macro area under the receiver operating characteristic curve (AUC) of 0.9975, with a mean accuracy of 93.72

± 1.06% across all the seeds. The attacks of white-box FGSM, PGD, CW-L2, and MI-FGSM greatly affected the accuracy, and the adversarial fine-tuned accuracy under FGSM attack was 74.17%, adversarial accuracy under PGD attack was less than 4% and adversarial accuracy under MI-FGSM attack was less than 4%, showing poor cross-attack generalization. A preliminary attack detector has been able to obtain an AUC of 0.857. Sometimes, clinically relevant regions were identified by Grad-CAM++, Score-CAM, and GradientSHAP, which also sometimes focused on non-anatomical artefacts. This means that achieving focused attention-augmenting and high accuracy on clean images does not necessarily mean that the accuracy of the adversarial images is also high, and that the explanation is good without adversarial training; it needs to be further verified by independent datasets and multi-center validation before applying to the clinical environment.

Keywords: Gastrointestinal endoscopy, MaxViT-Tiny, CBAM, Kvasir dataset, adversarial robustness, explainable artificial intelligence, image classification

1 Introduction

Gastrointestinal (GI) diseases include conditions that impact the esophagus, stomach, small intestine, and colon, such as esophagitis, ulcerative colitis, polyps, ulcers, and GI cancers. They undergo a high proportion of endoscopy to aid their diagnosis, and in the small bowel, they use wireless capsule endoscopy (WCE) to generate a large number of images that require interpretation by trained clinicians. Thus, Artificial Intelligence (AI) is attracting increasing attention, and recently, a systematic survey of the field showed that the number of publications on deep learning with Convolutional Neural Networks (CNNs) as the most popular deep learning architecture is steadily rising [1]. However, there are other use cases in which AI is used in tabular data analysis, for instance, the early detection of GI disorders [2], the anaesthetic during surgery [3], and measurement of slow waves in the stomach and duodenum [4], and many more.

Transfer Learning and Attention Mechanism (TLAM) and Transformer-based designs have recently been used in the area of Endoscopic Image Analysis (EIA). We have achieved 97.84% accuracy for an attention-guided variant of ResNet (SE + ResNet-H) for 5 different classes of HyperKvasir WCE images [5],focusedattention-augmenting and attention-augmentin transfer learning has been explored for 8 classes of the Kvasir [6] images. Some of the models that have been fine-tuned using meta-heuristics and tested on Kvasir include Hybrid rice optimization with attention-based LSTM [7] and improved ShuffleNet [8]. 98.2% has been reported on 8 classes from Kvasir [9], and 99.50% has been reported on the 4-class WCECCD dataset [10] through transformer-based architectures like TranfoNet. In addition to classification, deep models have also been used for lesion detection [11, 12] and lesion segmentation [13].

But there are still several issues that need to be addressed. In the review [1], there are still some open points concerning limited real-time effectiveness, difficulties in integration in the hospital, lack of interpretability, and lack of labelled data. The second practice is the evaluation practice: All the image-based studies reported so

far have results on a single public benchmark typically Kvasir or a variant of it, and split at the level [5, 8–10]. Unfortunately, when leakage-aware incorporation of group-level split and cross-dataset validation was done, it turned out that all the methods compared suffered, and that the difference with the numbers reported (97%-99%) might be too optimistic for the real world, where the improvement was limited to 2% on Kvasir-v2. Even some studies have suggested the need for multi-center or prospective validation [10, 14], as the performance of the fine-tuned models in transfer learning cannot be guaranteed over different devices, centersremoving duplicates, 3,996 images remained and were split into 3 sets: 70% for training, 15% for validation, and 15% for testing; we compared the model with 5 backbones using the same CBAM head across, or populations, and they would need to be validated on a l,arger and more diverse public dataset. It was also difficult to deploy: GIFCOS-DT was deployed on an embedded device running at around 5 frames per second (fps), whereas similar to 28 fps in the standard setting [12], and the models constructed from models in a clean benchmark data set [14] were impacted by illumination changes, blur, reflections and viewpoint variation. One important question not studied in the realm of GI imaging is adversarial vulnerability – small perturbations that are hard to detect, but can lead to prediction change of the deep networks in the medical imaging field, which is very important to clinical safety [15] [16].

These observations will be used as a basis on which to build an efficient, robust, and accurate, easily interpretable diagnosis model of GI. Lightweight architecture [1], explainable AI [1], and hybrid CNN-Transformer architecture [1] have been shown to have potential; there are improvements in incorporating attention mechanisms [5] and multimodal fusion [9], as well as the use of ensemble learning [7] and joint image quality screening and abnormality detection [12]. To support a model decision, such as by visualization as in the case of Grad-CAM [5, 9],multi-center has been tried, but not addressed in the studies on GI above. In general, none of the above issues of generalization and adversarial vulnerability are taken into account. Patient group splits (fierce testing) as well as cross-dataset testing and multi center validation are also of importance [10] and [14]. This study fills these gaps by using attention- and transformer-based representations, transparent and deduplicated seed-repeated evaluation and investigating adversarial robustness and explanation reliability.

This Work's Contributions

MaxViT-Tiny Classification Framework: We explore the performance of endoscopy image classification framework based on MaxViT-Tiny with the same classification head (Convolutional Block Attention Module (CBAM)) on eight classes of the Kvasir dataset when compared to five other backbone networks with the same classification head.

Comprehensive Performance Evaluation: The proposed model evaluation is based on accuracy, macro-precision, macro-recall, macro-F1 score and macro-AUC, and runs three times tests with different random seeds to test the model stability.

Adversarial Robustness Analysis: We study the adversarial robustness of the model against FGSM, PGD, CW-L2, and MI-FGSM attacks and adversarial distance, and then study the effect of adversarial fine-tuning on the classification robustness.

Explainability and Computational Analysis: Analyze the adversarial-image detection as well as computation cost in explaining the model prediction via explainability methods including Grad-CAM++, Score-CAM and GradientSHAP.

2 Literature Review

The rest of the studies (14) are analyzed by method and summarized in an overview (Table 1).

“‘latex

Table 1: Summary of the studies reviewed in Section 2 (NR = not reported in the extracted summary; ASR = attack success rate; MDL = minimum detection limit).

Ref.	Year	Dataset	Model/method	Validation	Key result
[15]	2024	Fundoscopy, Chest X-ray14, ISIC2019, Breast Histology	GAN attack on ViT	85/15 split	ASR 51.67–95.23%; clean acc. >85%
[16]	2023	MedMNIST-2D (12 datasets)	MedViT-T/S/L (ECB + feature permutation)	Compared with ResNet and others; FGSM/PGD	Higher acc. and AUC than baselines
[17]	2023	Kvasir-V2; 8 classes, 8,000 images	Improved CapsNet + class attention; MSSA tuning; DBN-ELM	Train/val curves; PR and ROC on test set	Outperforms compared methods (figures NR)
[18]	2022	Public WCE dataset, >47,000 images, imbalanced	RAt-CapsNet (volumetric attention + pyramid features)	NR	Acc. 98.51% (binary); >95.65% (multi-class)
[19]	2023	Kvasir v2; HyperKvasir	Swin-style ViT, hybrid shifted windows	HyperKvasir 60/20/20 split	Acc. 95.42% (Kvasir v2); 86.81% (HyperKvasir)
[20]	2023	Kvasir-Capsule, balanced top-9 classes	ViT, MobileNetV3Large, ResNet152V2; Grad-CAM, LIME, SHAP	Train/test; visual heatmap assessment	F1 97% ± 1% (ViT); others >90%
[21]	2022	HR-EGG, 32 human subjects	Robust regression + optimal-transport clustering	5 patients re-assessed after intervention	7 patients with normal slow waves but worse symptoms
[22]	2024	Pig GI fluids; PBS	pH-responsive MPO sensing capsule	Fluid vs. PBS readings; selectivity	<5% variation; sensitivity 6.14 μJ/(U/ml)
[23]	2024	142 studies screened, 47 included	Systematic review (PRISMA)	PRISMA screening	Needs: high gain, wideband, omnidirectional, low SAR
[24]	2024	451 exhaled breath samples	TDLAS + SVM	NR	MDL 42.12 ppbv; 100% on three subgroups
[25]	2026	VR-Caps gastric simulation	ASE-PPO (PPO + RND + spectral noise)	Domain randomization; ablations; p = 0.035	95.4% coverage; +4.15–6.17% vs. PPO
[26]	2026	OCT-c8, OCTID, UCSD-3	Pruned adversarially trained ViT-B/16	Six attacks; vs. lightweight ViTs	Acc. 89.20–94.78%; ∼45–50% lower FLOPs
[27]	2025	BreastMNIST, BloodMNIST, OrganCMNIST	ViT + HOG/LBP/SIFT + adversarial CNN	Compared with baseline ViT	Robust acc. +79% (FGSM), +77% (PGD)
[28]	2025	MedBIoT, N-BaIoT (network traffic)	MedViT + AdvTrans	Compared with baselines	Acc. 99.2%; F1 99.1%

“‘

In this section, models to classify Endoscopic images using Capsule Network and Transformer models are proposed.

With non-standard CNNs, there are a few studies to classify GI images. On Kvasir-V2 (8,000 images, eight classes), some modifications (median smoothing, CapsNet with class attention layer, salp swarm tuning, and DBN-ELM Classifier) were cited cite ref28, made by MSSADL-GITDC, and it was reported that it performed better compared to the compared methods, but no quantitative details were provided in the source summary [17]. In the binary classification case, the accuracy of RAt-CapsNet was 98.51%, and for the multi-class case, it was more than 95.65% (some pairs of

classes have few distinguishing features and are at low resolution), with more than 47,000 labeled images in the imbalanced class [18].

Transformer-based classification was explored in [19], where a Swin-style vision transformer with hybrid shifted windows was used, which achieved 95.42% accuracy on Kvasir v2 and 86.81% accuracy on HyperKvasir (with a 60/20/20 split), indicating that the accuracy is dependent on the difficulty level of the dataset. In [20], the authors use Grad-CAM, Grad-CAM++, LayerCAM, LIME, and SHAP to compare several pre-trained networks like a vision transformer (ViT), MobileNetV3Large, ResNet152V2, and other networks, on a balanced subset of the nine largest classes of Kvasir-Capsule for explainability purposes. The other two networks had an F1 score of more than 90% ± 1%, while the ViT achieved 97% ± 1%, but the explanation assessment was qualitative, and LIME considered black borders of images as important [20].

2.1 Physiological sensing and capsule hardware, robotic navigation.

The second group of studies is, based on either non-imaging signals or capsule hardware. Robust regression and optimal-transport clustering in a small cohort of 32 human subjects suggested that, although slow wave characteristics were normal, there was a non-myoelectric etiology in seven patients with worse symptoms [21]. In vivo results have not been reported, but a single-use ingestible capsule (11 mm × 26 mm) was developed that contained a chemiluminescent myeloperoxidase (MPO) sensor and had a pH-responsive coating which, when used in pig GI fluids to detect inflammation, showed less than 5% variation in reading and a sensitivity of 6.14 μJ/(U/ml) [22]. The outcome of the PRISMA evaluation of 47 of the studies is that the capsule-to-receiver telemetry links for WCE should be high gain, wide bandwidth, omnidirectional pattern, and have low specific absorption rate (SAR). However, if combined with SVM classification, TDLAS achieved 100% accuracy for ileus, ascending colon cancer, and cancers with liver metastasis, and no significant correlation was found for GIST, stomach cancer, or rectal cancer, which may be due to the small number of patients involved in the study[24]. In terms of autonomous capsule navigation, results achieved by ASE-PPO displayed 95.4% stomach coverage at 120 s of simulation, a 4.15%–6.17% improvement over PPO, but questions concerning the simulations and the real world, and the safety of the magnetic control still need to be answered [25].

Adversarial Robustness of Vision Transformers for Medical Imaging.

Finally, a last class of studies is about adversarial robustness, and transformers are beginning to be used in medical imaging. These works don't actually utilize GI information, but may be relevant to issues germane to endoscopic models. An attack generator that uses GANs has been used in an attack for Fundoscopy, Chest X-ray14, ISIC2019, and Breast histology data, with attack success rates of 81.36%, 95.23%, 77.75%, and 51.67%, respectively, with the lowest attack success rate from the smallest dataset [15]. GreenShield trained a ViT-B/16 and adversarially trained with anti-saturation in three retinal OCT dataset cases, saving 45% - 50% FLOPs, inference time, and carbon footprint, while maintaining higher robustness to six white-box attacks [26] compared to other lightweight ViTs. Hy-Deft improves the baseline ViT by up to 79% (FGSM) and 77% (PGD) in terms of robust accuracy on three MedMNIST

datasets [27] by combining the features of ViT with HOG, LBP, and SIFT descriptors, and leveraging an adversarially trained light CNN. One of MedViT's modules, MAViT, integrated with the adversarial-aware transformer that obtained 99.2% on MedBIoT and N-BaIoT

Summary of study and gap for research are presented.

The studies analyzed as a whole showed that the capsule-network and transformer models had high accuracy for the endoscopic and capsule endoscopy images, respectively [17] – [20], but it was difficult to compare each model's accuracy because several studies used different evaluation protocols, datasetcentre-cropped sizes, class distributions, etc. Other approaches to diagnosis include physiological sensing, breath analysis and capsule hardware, but these have been tried in limited numbers, animals, with collected fluids, and in simulation [21-22, 24-25]. Although studies have been conducted on testing powerful yet lightweight vision transformers on various images such as retinal, chest, dermoscopic, and histological images and MedMNIST images [15, 16, 26–28], no study has been performed on GI endoscopic images. The present work is inspired by this lack of knowledge of the behaviour of both classifiers (attention-based and transformer-based GI classifiers) under perturbations, under patient-level validation, and across datasets.

3 Methodology

3.1 Dataset

The experiments are carried out on the original Kvasir data set (version 1), which comprises gastrointestinal (GI) endoscopic images (4,000 images, 500 per class) [29]. This is not the 8,000-image Kvasir-v2 release used in [17] and [19]. This data set consists of eight classes: three anatomical landmarks (normal-z-line, normal-pylorus, normal-cecum), three pathological findings (esophagitis, polyps, ulcerative-colitis), and two classes of endoscopic procedures (dyed-lifted-polyps, dyed-resection-margins). All files were previously hashed with the SHA256 algorithm, and duplicate files (which are bit-by-bit equal) were removed before splitting; the script aborts if two equal files have different labels. Following this step, 3,996 images were left (500 per class, polyps and ulcerative colitis have 498 images, respectively).

3.2 Data partitioning

All images were split into training, validation and test sets with stratified sampling using a fixed split seed of 42 to ensure that class proportions are maintained and all models and training seeds are evaluated on the same 600 test images (Table 2). For checkpoint selection, early stopping, and learning-rate scheduling, the validation set was used, and the test set was only used for the final evaluation. This split is only at the image level: patient or procedure identifiers are not included in the dataset, and near-duplicate frames from the same examination cannot be excluded because the byte-identical frames have been removed. This is actually a restriction of the assessment protocol.

Table 2: Number of images per class in each subset.

Class	Train	Validation	Test	Total
dyed-lifted-polyps	350	75	75	500
dyed-resection-margins	350	75	75	500
esophagitis	350	75	75	500
normal-cecum	350	75	75	500
normal-pylorus	350	75	75	500
normal-z-line	350	75	75	500
polyps	348	75	75	498
ulcerative-colitis	349	74	75	498
Total	2797	599	600	3996

Test set: 75 images per class.

3.3 Preprocessing and data augmentation

All images were converted to RGB, resized to 256 × 256 pixels, and normalised with the ImageNet channel means (0.485, 0.456, 0.406) and standard deviations (0.229, 0.224, 0.225). On-the-fly augmentation of training images included a random resized crop (scale 0.80 – 1.0), random horizontal flip, colour jitter (brightness 0.15, contrast 0.15, saturation 0.10) and a random rotation (up to 12°). Validation and test images were only resized to 256 × 256 and centre-c,ropped to 224 × 224. Class imbalance was automatically verified: weighted random sampler is only enabled if the ratio between the largest and smallest training class is $\geq$1.25. The classes are almost balanced (ratio approx. 1.01); therefore, ordinary random shuffling of training images is used, and class-weighted cross-entropy is used. In Fig. 1, two representative images for each class are displayed.

3.4 Proposed model architecture

The proposed network is composed of two components: the MaxViT-Tiny backbone [30] and Convolutional Block Attention Module (CBAM) [31] (Fig. 2). All the backbones used in this study are compared using the same CBAM head that is attached to the final feature map of all the backbones. The MaxViT-Tiny is a hybrid convolution–transformer network. It is composed of a convolutional stem and 4 levels of MaxViT blocks, each of which is a stack of MBConv layers, block (windowed local) self-attention, and grid (dilated global) self-attention, enabling both local and global dependencies to be captured at linear complexity. The backbone was initialised from 'timm' library checkpoint maxvit_tiny_tf_224.in1k which was pretrained on the ImageNet-1k dataset and the original classification head of the backbone was removed, leaving the last feature map of the backbone (7 × 7 × 512 for 224 × 224 input) as the output.

$$M_c(F) = \sigma(W_1\ \mathrm{ReLU}(W_0\,\mathrm{GAP}(F)) + W_1\ \mathrm{ReLU}(W_0\,\mathrm{GMP}(F))) \tag{1}$$

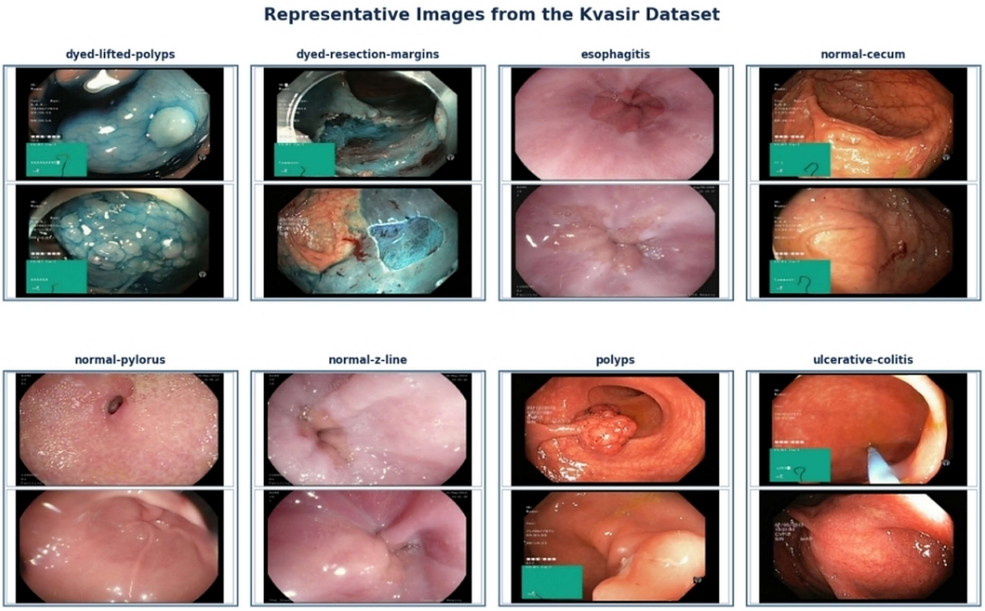


Fig. 1: Sample input images

$$F' = M_{\mathrm{c}}(F) \otimes F \tag{2}$$

The spatial-attention branch concatenates the channel-wise average and maximum maps of F′, applies a 7 × 7 convolution and a sigmoid gate, and rescales the spatial positions (Eqs. 3 and 4):

$$M_{\mathrm{s}}(F') = \sigma\big(f^{7\times7}([\mathrm{AvgPool}_{\mathrm{c}}(F'); \mathrm{MaxPool}_{\mathrm{c}}(F')])\big) \tag{3}$$

$$F'' = M_{\mathrm{s}}(F') \otimes F' \tag{4}$$

where σ is the sigmoid function, W0 and W1 are the weights of the shared perceptron, $\otimes$ is element-wise multiplication with broadcasting, and $[\cdot;\ \cdot]$ is concatenation along the channel axis. The refined feature map F″ is reduced by global average pooling, passed through dropout (p = 0.2) and a fully connected layer with eight outputs, followed by a softmax.

3.5 Training procedure

The model was trained using end-to-end automatic mixed precision with AdamW optimiser for a maximum of 30 epochs, $tildey_k\mathrm{n}_k, withabatchsizeof16andinitiallearningrateof2\times\ 10^{-4}$, and weight decay of 10^{-4}. The loss function used was cross-entropy with label smoothing [34] (Eq. 5) and class-weighted. The main settings are summariz$F'7\ times7$ed in Table 3.

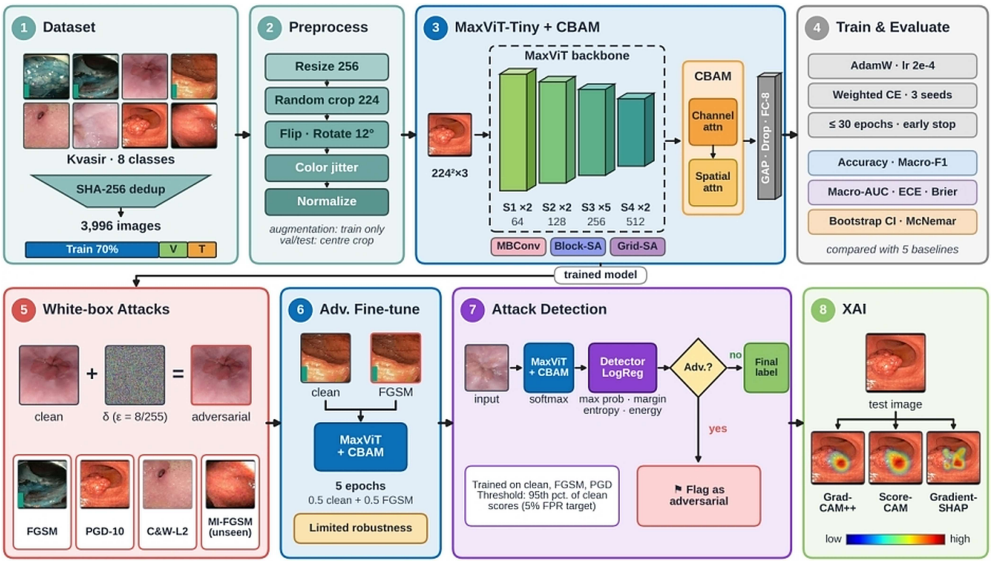


Fig. 2: Architecture of the proposed MaxViT-Tiny classifier with a CBAM head.

$$\mathcal{L}_{\mathrm{CE}} = -\sum_{k=1}^{K} w_k \, \tilde{y}_k \log p_k \tag{5}$$

where K = 8 represents the number of classes, pk represents the predicted probability of class k, $\tilde{y}$k is the smoothed label (yk is the one-hot labeled class, and α = 0.05), and wk represents the class weight (nk is the number of training images of class k and n is the total number of training images). ReduceLROnPlthe ateau: The learning rate was reduced by half if the validation loss did not improve for three epochs. After each training epoch, the checkpoint that achieved the highest validation macro-F1 score was retained, and training was terminated if the validation macro-F1 score did not improve for 8 consecutive epochs. In the reported run (seed 44), the best checkpoint was found at epoch 3 (validation macro-F1 0.9533) and training was terminated at epoch 11 (Fig. 3). Training was repeated with three random seeds (42,43 and 44) and the data split was unchanged. For reproducibility, deterministic cuDNN and the same seeds for Python, NumPy and PyTorch were used.

Table 3: Implementation details and hyperparameters.

Item	Setting
Backbone	MaxViT-Tiny (maxvit_tiny_tf_224.in1k), ImageNet-1k pretrained
Attention module	CBAM on final feature map (reduction 16, 7×7 spatial kernel)
Classifier head	Global average pooling, dropout 0.2, linear layer (8 classes)
Input size	224×224 (256×256 resize before crop)
Optimiser	AdamW, lr 2×10^{-4}, weight decay 10^{-4}
Batch size/max epochs	16 / 30
LR scheduler	ReduceLROnPlateau on validation loss (factor 0.5, patience 3)
Early stopping	Validation macro-F1, patience 8
Loss	Class-weighted cross-entropy, label smoothing 0.05
Class balancing	Random shuffle (imbalance ratio ≈ 1.01 ¡ 1.25)
Adversarial fine-tuning	5 epochs, lr 5×10^{-5}, FGSM ($\epsilon = 8/255$), 0.5 clean + 0.5 adversarial loss
Seeds	Training seeds 42, 43, 44; split seed 42
Hardware/software	NVIDIA Tesla T4; Python 3.13, PyTorch 2.11, torchvision 0.26, timm 1.0.30, scikit-learn 1.6.1

3.6 Adversarial fine-tuning

To improve robustness, the best clean checkpoint was fine-tuned for five epochs with a mixed clean and adversarial objective (Eq. 6), where the adversarial images were generated on the fly with an untargeted FGSM attack [35] (Eq. 7):

$$\mathcal{L}_{\mathrm{AT}} = 0.5\,\mathcal{L}_{\mathrm{CE}}(f_\theta(x), y) + 0.5\,\mathcal{L}_{\mathrm{CE}}(f_\theta(x_{\mathrm{adv}}), y) \tag{6}$$

$$x_{\mathrm{adv}} = \mathrm{clip}(x + \epsilon\ \mathrm{sign}(\nabla_x \mathcal{L}_{\mathrm{CE}}(f_\theta(x), y))) \tag{7}$$

with $\epsilon = 8/255$ in pixel space (applied in normalized space) and clip$(\cdot)$ keeping the image in the valid range. The learning rate was reduced to one quarter of the original value (5×10^{-5}), and the checkpoint with the highest clean validation macro-F1 was retained (Fig. 4).

3.7 Baseline models and comparison protocol

Five models were compared to the proposed model: ConvNeXt-Tiny [36] and a hybrid model of ConvNeXt-Tiny [36] and Swin-Tiny [37], DenseNet121 [38], GhostNet-100 [39], and MobileNetV3-Large-100 [40]. For all models, the results were tested on the same 600 test images, and the order of the images in the prediction files stored by the models was found to be the same; they used three different seeds (42, 43, 44) for training. The main comparison is given for seed 44; consistency across seeds is given as mean ± standard deviation. The same CBAM module is applied to the final feature map as the proposed model for all the baselines; hence, this is a backbone comparison. The class-balancing was different for each model: ConvNeXt-Tiny, DenseNet121, GhostNet, and MobileNetV3 used a weighted random sampler, while MaxViT-Tiny and ConvNeXt-Tiny + Swin-Tiny hybrid used ordinary random shuffling.

3.8 Evaluation metrics

Performance was measured on the test set with the following metrics. Let TP, TN, FP, and FN denote the one-vs-rest counts for each class, and let K = 8 be the number of classes.

Accuracy is the fraction of correctly classified test images. The metrics are defined as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN},$$
$$\text{Precision} = \frac{TP}{TP + FP}, \qquad \text{Recall} = \frac{TP}{TP + FN},$$
$$\text{Specificity} = \frac{TN}{TN + FP}, \qquad \text{F1} - \text{score} = \frac{2\,\text{Precision}\,\text{Recall}}{\text{Precision} + \text{Recall}}.$$

Class-wise values are reported together with macro averages, computed as the unweighted mean over classes, and weighted averages. For any metric M, the macro average is:

$$M_{\text{macro}} = \frac{1}{K}\sum_{k=1}^{K} M_k.$$
$$F1_{\text{macro}} = \frac{1}{K}\sum_{k=1}^{K} F1_k \tag{8}$$

Cohen's kappa [41], reported as a supplementary measure of agreement.

Area under the ROC curve (AUC), computed one-vs-rest for each class and macro-averaged.

Confusion matrix (raw counts and row-normalized) to analyze the errors between classes.

Statistical significance: McNemar's exact test [42] on the paired test predictions of two models, using the discordant pairs, where b is the number of images classified correctly only by the first model and c is the number classified correctly only by the second (Eq. 9). Each comparison was performed separately per seed (five baselines $\times$ three seeds = 15 tests), and two-sided exact p-values are reported. Bonferroni-corrected thresholds are given both within a seed ($0.05/5 = 0.01$) and across all 15 comparisons ($0.05/15 \approx 0.0033$), and the Holm step-down procedure is used as a less conservative check. Because the three seeds share the same split and test images, the tests are not independent replications, and McNemar's test addresses differences between the trained models on this test set only, not the variability due to training randomness or to a different patient sample.

$$p = \min\left(1,\ 2\sum_{i=0}^{\min(b,c)} \binom{b+c}{i}\left(\frac{1}{2}\right)^{b+c}\right) \tag{9}$$

Computational cost: number of parameters, model size, multiply–accumulate operations (MACs) and FLOPs for one 224 $\times$ 224 image (thop), and the inference latency

and throughput on the GPU for a single image (batch size 1) after 20 warm-up runs, averaged over 100 runs.

3.9 Adversarial robustness and attack detection

Robustness was evaluated on the complete test set with white-box, untargeted attacks against both the clean-trained and the adversarially fine-tuned checkpoint. FGSM followed Eq. (7). PGD [43] iterates the signed-gradient step with step size α and projects the result onto the ϵ-ball Bϵ(x) around the input (Eq. 10):

$$x^{t+1} = \Pi_{B_\epsilon(x)}\big(x^t + \alpha\ \mathrm{sign}\big(\nabla_x \mathcal{L}_{\mathrm{CE}}(f_\theta(x^t), y)\big)\big) \tag{10}$$

MI-FGSM [44] accumulates the L1-normalised gradient with decay factor μ (Eq. 11) and updates the image with the sign of the accumulated momentum (Eq. 12):

$$g^{t+1} = \mu g^t + \frac{\nabla_x \mathcal{L}_{\mathrm{CE}}(f_\theta(x), y)}{\|\nabla_x \mathcal{L}_{\mathrm{CE}}(f_\theta(x), y)\|_1} \tag{11}$$

$$x^{t+1} = \Pi_{B_\epsilon(x)}\big(x^t + \alpha\ \mathrm{sign}(g^{t+1})\big) \tag{12}$$

The CW-L2 attack [45] minimises the perturbation norm plus a margin loss on the logits Z, with confidence κ (Eq. 13):

$$\min_\delta\ \|\delta\|_2^2 + c\max\Big(\max_{j\neq y} Z_j(x+\delta) - Z_y(x+\delta),\ -\kappa\Big) \tag{13}$$

FGSM, PGD, and MI-FGSM used $\epsilon = 8/255$ in normalised pixel space; PGD used 10 steps ($\alpha = 2/255$) and a random start, while MI-FGSM used 10 steps ($\alpha = 2/255$) and $\mu = 0.9$ and was not used during fine-tuning. CW-L2 used 30 optimisation steps, c = 1, and $\kappa = 0$. We report the accuracy under attack (robust accuracy) and the attack success rate (ASR), defined as the fraction of originally correct predictions that were flipped by the attack (Eq. 14):

$$\mathrm{ASR} = \frac{\big|\{i \in S : f_\theta(x_i^{\mathrm{adv}}) \neq y_i\}\big|}{|S|} \tag{14}$$

where S is the set of test images correctly classified before the attack.

As a preliminary defence, a logistic-regression attack detector was trained on four statistics of the softmax output, computed from clean, FGSM, and PGD images drawn from 50 validation images per class (Eq. 15):

$$s(x) = \sigma\big(w^\top \phi(x) + b\big), \qquad \phi(x) = [\max_k p_k,\ p_{(1)} - p_{(2)},\ H(p),\ E(z)] \tag{15}$$

where p is the softmax vector, p(1) and p(2) are its two largest entries, z the logits, and the entropy H and energy score E [46] are given by Eq. (16):

$$H(p) = -\sum_{k=1}^{K} p_k \log p_k, \qquad E(z) = -\log\left(\sum_{k=1}^{K} e^{z_k}\right) \tag{16}$$

An image was flagged as adversarial when s(x) exceeded a threshold τ, set to the 95th percentile of the scores of 80 clean validation images (10 per class), targeting a 5% false-positive rate. The detector was evaluated on the clean and attacked test images.

3.10 Explainability

The model decisions were examined using three post hoc methods: Grad-CAM++ [47] on the final convolutional layer, Score-CAM [48] on the previous higher-resolution convolutional layer, and GradientSHAP [49] (Captum [50] zero and low-noise Gaussian baselines, 20 samples, noise std dev 0.09). The first correctly classified test image was visualised for each class, and the explanation was computed for the predicted class. These maps are of a qualitative nature.

3.11 External validation dataset

To check the generalisation beyond Kvasir, the trained model was also tested with an external dataset, GastroVision [51]. GastroVision is an open-access gastrointestinal endoscopy data set containing 8,000 images of 27 classes, including anatomical landmarks, pathological findings, polyp-removal cases, and normal cases. Images were collected at Bærum Hospital (Norway) and Karolinska University Hospital (Sweden) and labeled and reviewed by experienced endoscopists. The data were downloaded from the Kaggle mirror "orvile/gastrovision-gastrointestinal-disease-detection" that is hosted on the Open Science Framework (https://osf.io/84e7f/). No GastroVision image was used for training, validation, hyperparameter selection, or checkpoint selection, and the model was only trained on Kvasir.

The label sets for GastroVision and Kvasir are different. Only classes matching clinically the same Kvasir label were kept, and matching was performed on GastroVision folder name (Table 11). There are seven different counterparts to the eight classes of Kvasir. Those without a Kvasir equivalent class were not included (e.g., gastric polyps, resected polyps, and mucosal inflammation of the large bowel), and ulcerative colitis, which has none, was excluded. No re-check of mapping by a clinician.

All the selected files were subjected to hashing using SHA256 before the evaluation. A single repeated image in the external data was removed (byte-identical), and no external repeated image that was shared with Kvasir was found, resulting in 2,149 external images. Classes are imbalanced (819 vs 107). The data from Kvasir was collected at Vestre Viken Health Trust, including Bærum Hospital [29], and no patient identifiers are available for either of these. Therefore, near-duplicate frames or patient overlap between Kvasir and GastroVision cannot be excluded, and the evaluation is a cross-dataset test and not a fully independent multi-centre validation.

3.12 Preprocessing and protocol for the external evaluation

As with the Kvasir validation and test images (Section 3.3), the images from external sources were preprocessed as follows: converted to RGB, resized to 256 $\times$ 256, centre-cropped to 224 $\times$ 224, and normalised using the ImageNet means and standard deviations, but without augmentation. The seed-44 MaxViT-Tiny checkpoint was used in inference mode, with no weights adjusted, with a batch size of 32, and trained on Kvasir only. Since the model has eight outputs and the external data have seven classes, the softmax probabilities p were renormalized over the set S of the seven classes (Eq. 17), and all the metrics were computed on this seven-class problem, using the predicted class as the arg-max of the renormalized probabilities:

$$\tilde{p}_k = \frac{p_k}{\sum_{j \in S} p_j}, \qquad k \in S \tag{17}$$

The more stringent eight-way arg-max was also assessed, where a prediction of ulcerative colitis is considered an error. As in Section 3.8, accuracy, balanced accuracy, macro precision and recall, macro AUC (one-vs-rest) and MCC [52] and Cohen's kappa and ECE [53] (15 bins) were calculated. 2,000 class-stratified bootstrap resamples were used to find the 95% confidence interval for the macro F1-score. The results on the Kvasir test set were then compared with the external results in one table (Table 4).

4 Results and Discussion

4.1 Comparison with baseline models

Table 4 shows the comparison of the proposed MaxViT-Tiny with five baseline networks on the held-out test set (seed 44). MaxViT-Tiny exhibited the best results on all metrics: an accuracy of 94.33%, a macro F1-score of 0.9431, and a macro AUC score of 0.9975. The nearest baseline was the hybrid ConvNeXt-Tiny + Swin-Tiny (93.33%) with a difference of 1.00 percentage point. The other baselines (ConvNeXt-Tiny, DenseNet121, GhostNet, and MobileNetV3) were 3.50 – 12.16 percentage points lower than the proposed model. Macro-averaging is used to compute Precision, Recall, and F1-score for the eight classes.

Table 4: Performance comparison of the proposed MaxViT-Tiny model and baseline models on the test set (seed 44).

Model	Accuracy (%)	Precision	Recall	F1-score	Kappa	AUC
MobileNetV3	82.17	0.8392	0.8217	0.8196	0.7962	0.9850
GhostNet	89.67	0.8980	0.8967	0.8959	0.8819	0.9938
ConvNeXt-Tiny	90.83	0.9085	0.9083	0.9082	0.8952	0.9937
DenseNet121	90.83	0.9125	0.9083	0.9076	0.8952	0.9927
ConvNeXt-Tiny + Swin-Tiny	93.33	0.9343	0.9333	0.9332	0.9238	0.9942
MaxViT-Tiny (Proposed)	94.33	0.9459	0.9433	0.9431	0.9352	0.9975

4.2 Class-wise performance of the proposed model

The class-wise results of the proposed model are shown in Table 5. The AUC values of normal-cecum, normal-pylorus, and ulcerative-colitis classes were almost perfect (0.9934, 0.9868, and 0.9801, respectively), and each of these classes had an AUC value of at least 0.9924. The weakest classes were esophagitis (F1 = 0.8777) and normal-z-line (F1 = 0.8944). These two classes are mainly confused with each other: 14 of

the 75 esophagitis images were predicted as normal-z-line, which lowered the recall of esophagitis (0.8133) and the precision of normal-z-line (0.8372). Another source of error was the group of "dyed" images, where two dyed-resection-margins images were misclassified as "dyed-lifted-polyps" and vice versa.

Table 5: Class-wise results of the proposed MaxViT-Tiny (seed 44).

Class	Precision	Recall	F1-score	Specificity	AUC	Support
dyed-lifted-polyps	0.8987	0.9467	0.9221	0.9848	0.9974	75
dyed-resection-margins	0.9437	0.8933	0.9178	0.9924	0.9975	75
esophagitis	0.9531	0.8133	0.8777	0.9943	0.9928	75
normal-cecum	0.9868	1.0000	0.9934	0.9981	1.0000	75
normal-pylorus	0.9740	1.0000	0.9868	0.9962	1.0000	75
normal-z-line	0.8372	0.9600	0.8944	0.9733	0.9924	75
polyps	1.0000	0.9467	0.9726	1.0000	0.9998	75
ulcerative-colitis	0.9737	0.9867	0.9801	0.9962	0.9998	75
Macro average	0.9459	0.9433	0.9431	0.9919	0.9975	600

Specificity is computed one-vs-rest. Sensitivity equals recall and is therefore not repeated.

4.3 Training behaviour

The following is the training of the proposed model (seed 44) as illustrated in Fig. 3. The training loss decreased from 0.93 to 0.31 within 11 epochs, while the training accuracy increased from 74.6% to 98.7% and the validation loss stayed in the range of 0.42 to 0.49 with the validation accuracy ranging between 92% and 95%. The best validation macro-F1 score achieved (0.9533) was at the 3rd epoch, while early-stopping training stopped at the 11th epoch. The difference between training and validation accuracy suggests that there is some overfitting, which is to be expected since we are just fine-tuning a pretrained transformer on ˜2800 images only; it is for this reason that the checkpoint was saved at the validation step and not at the training step. The adversarial fine-tuning stage resulted in a very gradual decrease in the robust training loss (as seen in Fig. 4) over the five epochs without any noticeable drop in the clean validation loss (zoomed along the vertical axis in Fig. 4) on clean validation images.

4.4 Error analysis

Fig. 5 shows the confusion matrix of the proposed MaxViT-Tiny model, which is row-normalized. The model was able to obtain a recall of at least 0.81 for each of the eight classes and a mean per-class recall of 0.9433, which was a good recall for

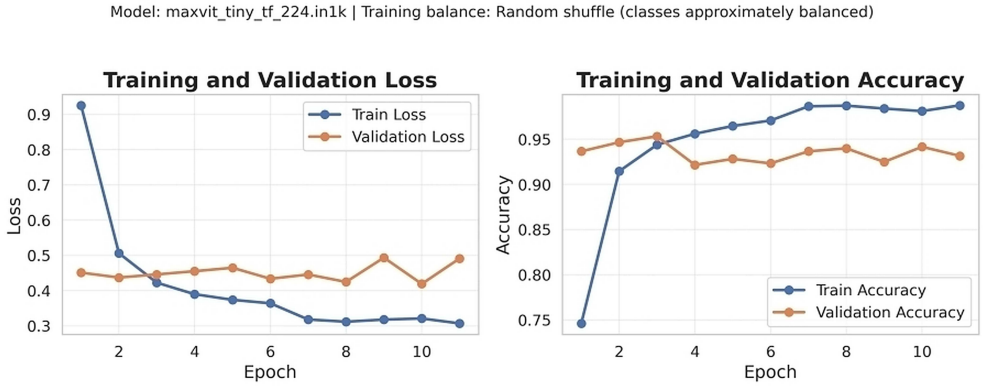


Fig. 3: Training and validation loss and accuracy of the proposed model (seed 44).

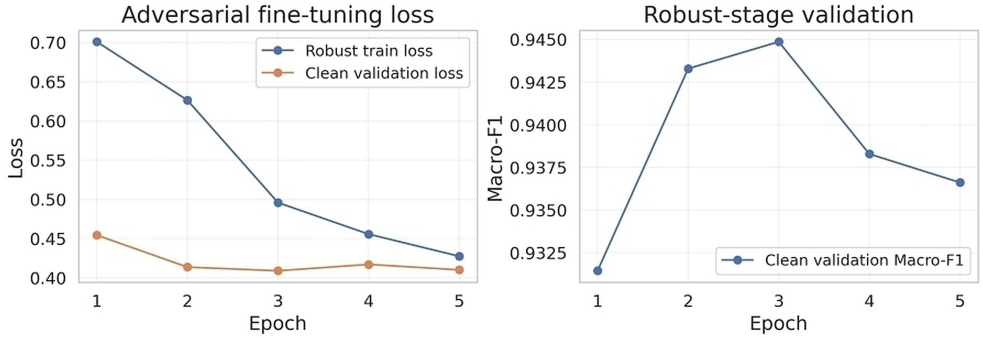


Fig. 4: Adversarial fine-tuning: training loss, clean validation loss, and clean validation macro-F1 per epoch.

all anatomical landmarks, pathological findings, and endoscopic procedures. Perfect recall (1.00) was obtained for normal-cecum, normal-pylorus, ulcerative-colitis (0.99), normal-z-line (0.96), and dyed-lifted-polyps (0.95). Misclassifications did not occur at random but tended to be grouped in clinically and visually close pairs of classes. Esophagitis had the lowest recall (0.81), with 19% of samples being classified as normal-z-line, a figure that is likely explained by the similar appearance of the gastric and oesophageal mucosa in most cases where early inflammatory changes may be subtle at the gastroes,ophageal (GE) junction. Likewise, dyed-resection-margins (0.89) and dyed-lifted-polyps (0.95) were confused in the two directions, with 11% and 5% of the samples misassigned, respectively; this may be explained by the similar chromoendoscopic dye-staining properties. All other off-diagonal entries were no more than 0.04, indicating that there was not much confusion between classes at the other locations.

One-vs-rest receiver operating characteristic (ROC) curves for each of the eight classes are shown in Fig. 6. The discrimination performance of the model is very high,

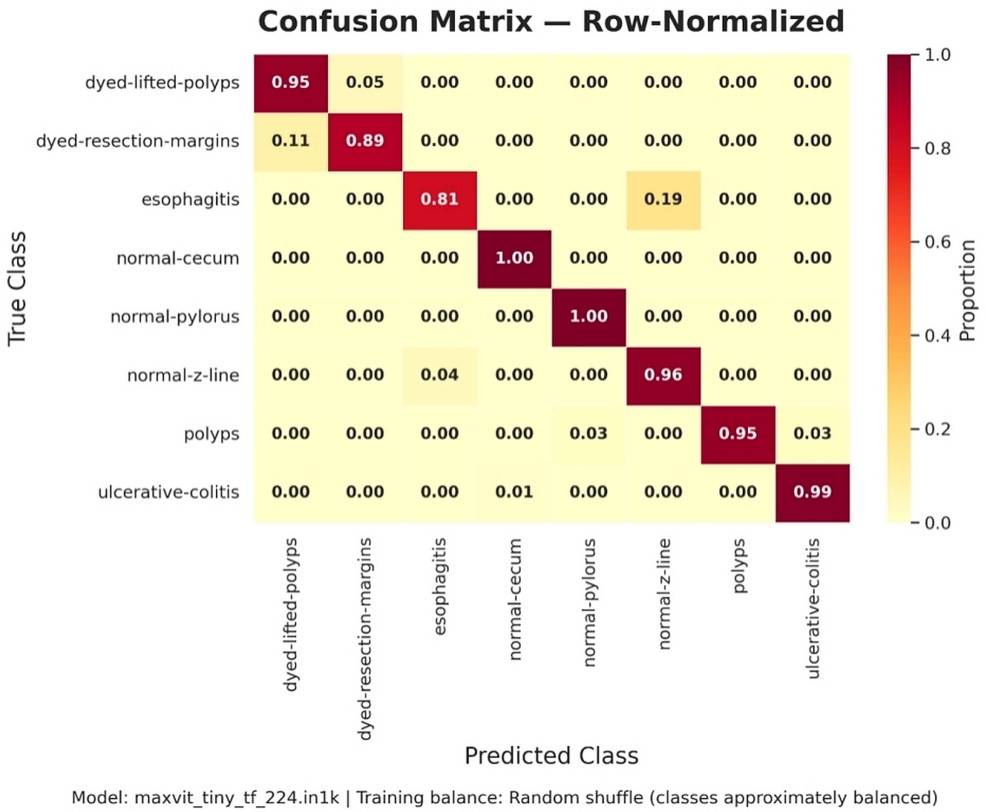


Fig. 5: Confusion matrix of the proposed model on the test set (seed 44).

with 0.992— 1.000 per-class AUC and 0.9975 macro-averaged AUC. Near-perfect separability (AUCs ≥ 0.9998) was achieved for normal-cecum, normal-pylorus, polyps, and ulcerative-colitis, while dyed-lifted polyps and dyed-resection regions yielded AUCs of 0.9974 and 0.9975, respectively. The lowest values occurred for the classes esophagitis (0.993) and normal-z-line (0.992), which are the two classes that were primarily confused with each other, as shown in the mutual confusion pattern in Fig. 5. Of particular interest, all of the curves showed an abrupt ascent to the upper left corner, with esophagitis having slightly better curves, and a true positive rate of greater than 0.9 at a false positive rate of less than 0.05 for most of the classes. These results suggest that the final error is largely due to visually similar classes and due to the fact that the model is not able to learn features that separate them, and not because of the lack of features, as the threshold is set in the classification process and is dependent on the classes.

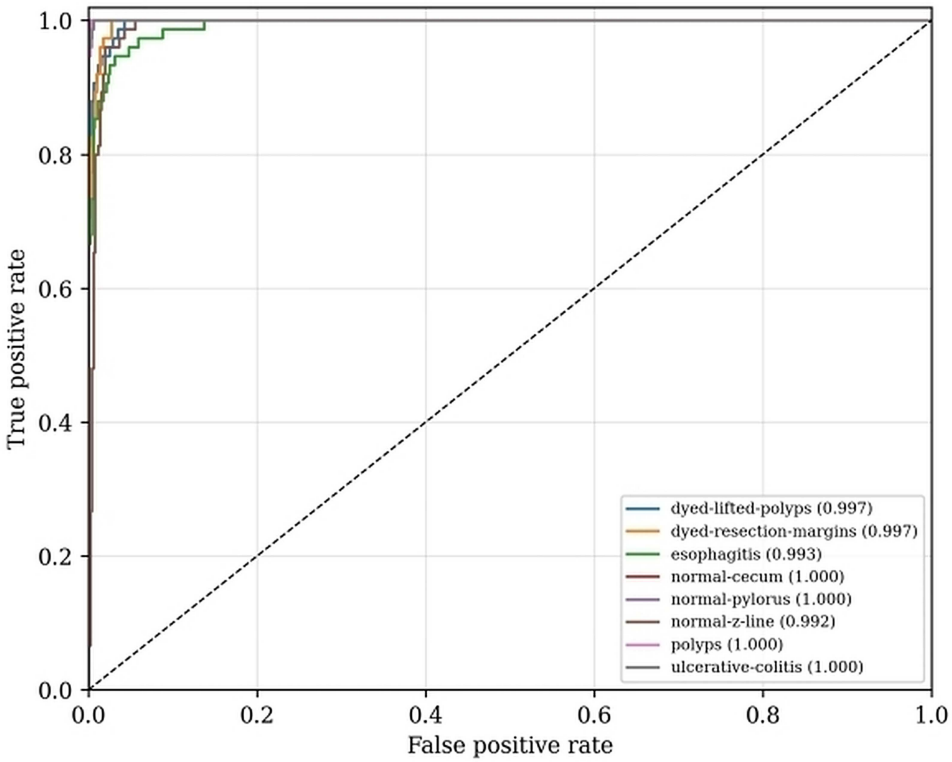


Fig. 6: One-vs-rest ROC curves of the proposed model (AUC in the legend).

4.5 Statistical significance analysis

McNemar's exact test was performed on the paired predictions on the same 600 test images for each of the three seeds (42, 43, and 44) to test whether the difference between the proposed model and each baseline model is statistically significant. Table 6 includes the number of images for which only the proposed model was correct (b), baseline only (c), and the two-sided p-value for each baseline.

Table 6: McNemar's test: proposed MaxViT-Tiny versus baseline models for seeds 42, 43, and 44.

Baseline	Seed 42: b / c	p-value	Seed 43: b / c	p-value	Seed 44: b / c	p-value
ConvNeXt-Tiny	23 / 19	0.644	29 / 9	0.0017**	31 / 10	0.0015**
ConvNeXt-Tiny + Swin-Tiny	18 / 20	0.871	21 / 12	0.163	18 / 12	0.362
DenseNet121	15 / 16	1.000	23 / 11	0.058	33 / 12	0.0025**
GhostNet	22 / 14	0.243	28 / 10	0.0051**	38 / 10	¡0.001***
MobileNetV3	74 / 16	¡0.001***	70 / 17	¡0.001***	84 / 11	¡0.001***

The proposed model showed a significant improvement in all three seeds (p ¡ 0.001) compared to MobileNetV3. In seeds 43 and 44, it was also significantly better than ConvNeXt-Tiny (p = 0.0017 and 0.0015) and GhostNet (p = 0.0051 and p ¡ 0.001), and in seed 44, it was additionally better than DenseNet121 (p = 0.0025). All of these differences are still significant at the per-seed Bonferroni criterion of 0.01. The Bonferroni threshold is $0.05/15 \approx 0.0033$, and all differences remain significant except GhostNet in seed 43 (p = 0.0051); using the less conservative Holm procedure, all 8 significant differences remain. However, the significance was not the same for all the seeds: in seed 42 only MobileNetV3 was significantly outperformed by the proposed model, while there was no significant difference in the other models, namely ConvNeXt-Tiny + Swin-Tiny (18 versus 20 discordant images) and DenseNet121 (15 versus 16). No differences were significant against the hybrid in any seed ($p \geq 0.163$), and thus this model can be considered statistically equivalent to the proposed model. Two limitations apply. Because the two trained models are compared on a fixed 600-image test set, McNemar's test indicates whether the models are different or not on these images, but not the amount of difference that can be attributed to randomness of training or the difference in patient population. Also, all three seeds have the same split and test images, and the three sets of tests are not independent replicates, and therefore were not combined to form a single p-value. Results thus suggest a consistent improvement over MobileNetV3 and a seed-dependent improvement over ConvNeXt-Tiny, GhostNet and DenseNet121.

4.6 Consistency across seeds

The accuracy over the three seeds is summarized in Table 7. The proposed MaxViT-Tiny model achieved the highest mean accuracy ($93.72 \pm 1.06\%$) and mean F1-score (0.9370 ± 0.0108) among all the models, followed by the ConvNeXt-Tiny + Swin-Tiny hybrid ($93.00 \pm 0.29\%$). The lower seed 42 (92.50%) compared to the other seeds 43 and 44 (94.33% in both seeds) is responsible for the larger standard deviation in the proposed MaxViT-Tiny model.

Table 7: Test accuracy (%) per seed and mean ± standard deviation across seeds.

Model	Seed 42	Seed 43	Seed 44	Acc. mean ± SD (%)	F1 mean ± SD
ConvNeXt-Tiny	91.83	91.00	90.83	91.22 ± 0.54	0.9120 ± 0.0054
ConvNeXt-Tiny + Swin-Tiny	92.83	92.83	93.33	93.00 ± 0.29	0.9298 ± 0.0030
DenseNet121	92.67	92.33	90.83	91.94 ± 0.98	0.9189 ± 0.0100
GhostNet	91.17	91.33	89.67	90.72 ± 0.92	0.9064 ± 0.0092
MobileNetV3	82.83	85.50	82.17	83.50 ± 1.76	0.8333 ± 0.0185
MaxViT-Tiny (Proposed)	92.50	94.33	94.33	93.72 ± 1.06	0.9370 ± 0.0108

Standard deviation is the sample SD (n = 3 seeds).

4.7 Computational cost

The proposed MaxViT-Tiny model has 30.44 M parameters (all trainable), a size of 116.12 MB, and requires 5.33 G MACs (10.66 GFLOPs) for a 224 × 224 image (Table 8). On a Tesla T4 GPU, the inference latency was 32.59 ms per image (30.69 images per second) at batch size 1, which is suitable for offline analysis, but it is not a lightweight network compared with mobile architectures such as MobileNetV3 and GhostNet.

Table 8: Computational complexity of the proposed model.

Measure	Value
Input size	224 × 224
Parameters (total/trainable)	30.44 M / 30.44 M
Model size	116.12 MB
MACs	5.33 G
FLOPs	10.66 G
Inference latency (batch size 1)	32.59 ms/image
Throughput	30.69 images/s
Device	NVIDIA Tesla T4

4.8 Robustness to Adversarial Attacks

The results of the proposed MaxViT-Tiny model when subjected to adversarial attacks are presented in Table 9. The series of attacks resulted in a major effect on both checkpoints. The model trained with clean data achieved an accuracy of 94.33% against

the Projected Gradient Descent (PGD) attack, 4.50% against the Carlini–Wagner L2 (CW-L2) attack, and 1.17% against the Momentum Iterative Fast Gradient Sign Method (MI-FGSM). According to the Fast Gradient Sign Method (FGSM), it could achieve a solid accuracy of 67.17% on the model that is trained on clean data.

Using adversarial fine-tuning, the accuracy of the model under FGSM is improved from 67.17% to 74.17%, and under CW-L2 is improved from 4.50% to 12.00%. Even the accuracy of the model under PGD and MI-FGSM was not high, at 3.33% and 3.00%, respectively. These results suggest that a single-step attack can only provide limited protection from stronger iterative attacks. The adversarial fine-tuned model had a clean accuracy of 92.50%, while the clean-trained model had an accuracy of 94.33% in clean image classification, suggesting that the adversarial fine-tuned model had a minor decline in its clean image classification performance.

Overall, the results show that while the proposed model is capable of performing well on clean images, it is still susceptible to adversarial perturbations. Adversarial fine-tuning with FGSM is proven to be effective in enhancing robustness against FGSM and CW-L2 but fails to defend against PGD and the unseen attack (MI-FGSM). As a consequence, the proposed model should not be considered to be adversarially robust without exploring other more robust defence approaches, e.g., multi-step adversarial training.

Table 9: Clean and adversarial accuracy on the test set (600 images).

Checkpoint	Attack	Clean Accuracy (%)	Robust Accuracy (%)	ASR (%)
Clean-trained	FGSM	94.33	67.17	28.80
Clean-trained	PGD	94.33	0.33	99.65
Clean-trained	CW-L2	94.33	4.50	95.23
Clean-trained	MI-FGSM (unseen)	94.33	1.17	98.76
Adversarially fine-tuned	FGSM	92.50	74.17	20.00
Adversarially fine-tuned	PGD	92.50	3.33	96.40
Adversarially fine-tuned	CW-L2	92.50	12.00	87.03
Adversarially fine-tuned	MI-FGSM (unseen)	92.50	3.00	96.76

4.9 Attack detection

The accuracy of the attack detector on the clean test images and the images generated by the four attacks is 64.1% with an AUC of 0.857 (Table 10). After thresholding on validation data, the false positive rate for clean test images was 4.3%, precision was 98.1%, and the success rate for attacked test images was only 56.2%. It could detect PGD with 83.8%, MI-FGSM with 91.0%, and FGSM and CW-L2 with 26.5% and 23.8%, respectively. The detector is therefore a preliminary part of the defence only and not necessarily a complete defence.

Table 10: Performance of the attack detector on the test set.

Accuracy	Precision	Recall	F1-score	AUC	Balanced acc.
0.6410	0.9811	0.5621	0.7147	0.8573	0.7594

4.10 Qualitative explanation

The one test image corresponding to each of the 8 classes that were correctly classified (with confidence of 0.854-0.980) is shown with the corresponding explanation maps in Figure 7. The maps used in several classes are based on the lesion or on the anatomical landmark. For dyed-resection margins, both Grad-CAM and Score-CAM target the resection margin at the edge of the dyed area; in the case of normal-pylorus, both target the pyloric opening; and for polyps, Score-CAM covers the polyp mass. For esophagitis, Score-CAM concentrates on the mucosa next to the lumen, while the Grad-CAM++ map is more scattered, with another peak to the left side of the image. In the case of dyed-lifted polyps, Grad-CAM++ can segment the whole mucosal fold and dyed area, whereas Score-CAM can segment poorly and sparsely. In other examples, however, the maps highlight non-anatomical areas: Grad-CAM++ highlights the on-screen annotation box in normal-cecum and polyps, and the image border in normal-z-line, as well as the time-stamp area in ulcerative-colitis, and Score-CAM also weights text overlays in normal-z-line and ulcerative-colitis. This suggests the network may use these image artefacts to a certain extent, which are present in some of the Kvasir images. The graph does not indicate the number of times this occurs, and these maps are only qualitative evidence, not clinical validation - only one picture per class is displayed. GradientSHAP generated sparse pixel-level attributions which are harder to understand visually.

4.11 External validation on GastroVision

Table 11 lists the seven GastroVision classes that were retained, the Kvasir classes they were mapped to, and the number of external images per class.

Table 11: Class mapping between GastroVision and Kvasir and the number of external images used per class.

GastroVision folder	Kvasir class	Images used
Cecum	normal-cecum	113
Colon polyps	polyps	819
Dyed-lifted-polyps	dyed-lifted-polyps	141
Dyed-resection-margins	dyed-resection-margins	246
Esophagitis	esophagitis	107
Gastroesophageal_junction_normal z-line	normal-z-line	330
Pylorus	normal-pylorus	393
Total	7 classes	2,149

The results of the proposed MaxViT-Tiny model on the Kvasir test set (600 images, eight classes) and the external set of GastroVision images (2,149 images, seven shared

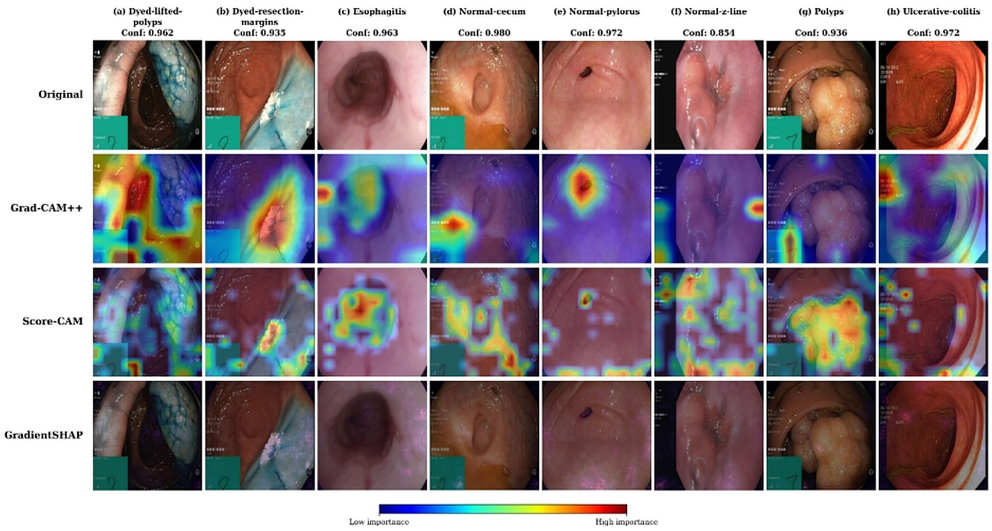


Fig. 7: Explanation maps (Grad-CAM++, Score-CAM, GradientSHAP) for one correctly classified test image of each of the eight classes (seed 44). For each row, the true class, the predicted class, and the prediction confidence are given on the left.

classes) were compared in Table 12. On GastroVision, the model reached an accuracy of 86.23%, a macro F1-score of 0.8025 (95% CI 0.7832 to 0.8233), a Cohen's kappa of 0.8220, and a macro AUC of 0.9821, compared with 94.33%, 0.9431, 0.9352, and 0.9975 on Kvasir. The accuracy consequently dropped by 8.10 percentage points, the macro F1-score dropped by 14.06 percentage points, and the AUC dropped 0.0154 only. The two evaluations are not exactly like-for-like since the Kvasir test set has eight balanced classes, while the external one has seven classes and is imbalanced. The decrease in macro F1 score is more severe than in accuracy since the smaller external classes are not identified to the same extent (Table 13). The highest accuracy, in the eight-way evaluation, was 82.18%, and 5.30% of external images were classified as the ulcerative-colitis class (which is not included in the external data).

Table 12: Performance of the proposed model on the Kvasir test set and on the external GastroVision dataset (seed 44).

Dataset	Acc. (%)	Precision	Recall	F1-score	Kappa	AUC
Kvasir (test set; 600 images, 8 classes)	94.33	0.9459	0.9433	0.9431	0.9352	0.9975
GastroVision (external; 2,149 images, 7 classes)	86.23	0.8213	0.7982	0.8025	0.8220	0.9821
Difference (external − Kvasir)	−8.10	−0.1246	−0.1451	−0.1406	−0.1132	−0.0154

Precision, recall, and F1-score are macro-averaged; macro recall equals balanced accuracy. The Kvasir values are taken from Table 4. The GastroVision values are computed over the seven shared classes with the softmax renormalised over these classes (Eq. 17).

Table 13: Class-wise results of the proposed model on the GastroVision external set (seed 44).

Class	Support	Precision	Recall	F1-score	AUC
dyed-lifted-polyps	141	0.7834	0.8723	0.8255	0.9852
dyed-resection-margins	246	0.9372	0.9106	0.9237	0.9962
esophagitis	107	0.4662	0.5794	0.5167	0.9622
normal-cecum	113	0.9351	0.6372	0.7579	0.9766
normal-pylorus	393	0.9625	0.7837	0.8640	0.9892
normal-z-line	330	0.7323	0.8455	0.7848	0.9754
polyps	819	0.9323	0.9585	0.9452	0.9900
Macro average	2,149	0.8213	0.7982	0.8025	0.9821

The row-normalized confusion matrix of the proposed model is shown in Fig. 8. Polyps (0.96), dyed-resection-margins (0.91), and dyed-lifted-polyps (0.87) were recognised best. The same pair of classes that caused most errors on Kvasir (Section 4.4) is the main problem on the external data as well: the weakest class was esophagitis (F1-score 0.517, precision 0.466), with 43 out of 107 images (40%) being predicted as normal-z-line, and 50 out of 330 normal-z-line images (15%) being predicted as esophagitis. The precision of normal-cecum (0.935) was high due to the low proportion of false positives (30%, 34/113), while the recall (0.637) was low since 14% (56/393)

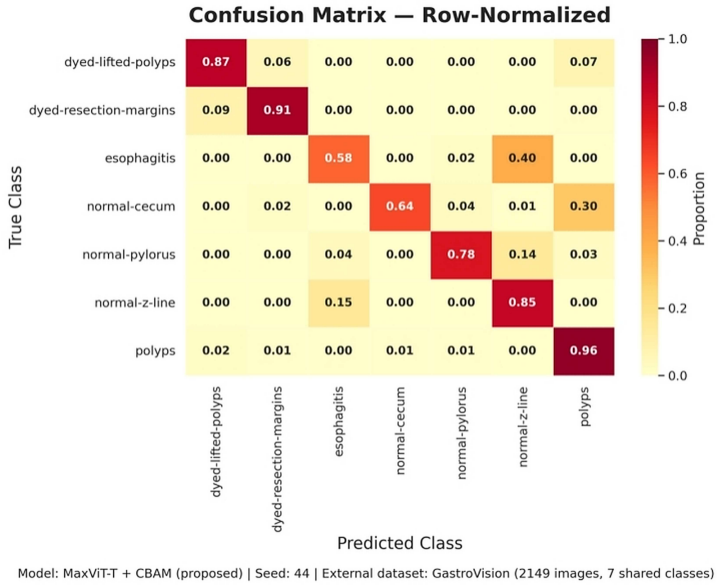


Fig. 8: Row-normalized confusion matrix of the proposed model on the GastroVision external set (seed 44, 2,149 images, 7 shared classes).

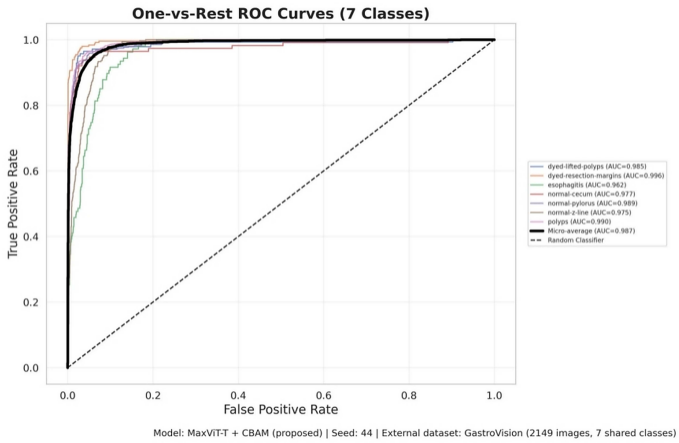


Fig. 9: One-vs-rest ROC curves of the proposed model on the GastroVision external set (AUC in the legend).

of the normal-pylorus were predicted as normal-z-line. ROC curves for all classes were high (Fig. 9), with AUC between 0.962 and 0.996. This means that the ranking of the classes is well-kept, and the decision boundaries learned on Kvasir do not fit the new data as well, which was not tested here. The ECE on GastroVision was 0.039.

There are several limitations to this external evaluation. One training seed was tested; external results may differ when using other seeds. Only 7 of 8 classes of Kvasir

could be examined, and the class mapping was done based on the name of the folders; it was assumed that ulcerative colitis was not included. The test is not completely source-institution independent since both Kvasir and GastroVision were partially collected at Bærum Hospital. Finally, the external classes are not well-balanced with the smallest classes with 107 and 113 images respectively and class-wise scores are not certain. The results show that the performance of the model on an independent test set is still good, although the performance drops significantly, and does not in itself demonstrate the robustness of the model across centres or devices.

5 Discussion

The proposed MaxViT-Tiny model outperformed all five baselines in terms of all the metrics (94.33% accuracy, macro F1 of 0.9431, macro AUC of 0.9975) in the seed-44 comparison, which validated the applicability of a hybrid convolution–transformer backbone for endoscopic classification. It did not show a significant advantage in any seed and was only 1.00 percentage point (pp) (or 6 images) better than its best baseline, ConvNeXt-Tiny + Swin-Tiny hybrid, and was significantly better than MobileNetV3 in all seeds but significantly better than ConvNeXt-Tiny, GhostNet, and DenseNet121 in only some seeds. While duplicate files have been removed from all of the models and each one was tested on the same test set, the accuracy is not as high as that of previous studies on Kvasir variants and capsule-endoscopy datasets [5] (97%) and [9] (99%) that often use random image-level partitions, which tend to overestimate generalization [14]. In the case of errors, they occurred around a pair of visually similar classes (esophagitis, normal-z-line), and some of them were due to threshold-based decisions; per-class AUC values were at or above 0.992, indicating that the classes are relatively distinguishable. More importantly, high clean accuracy was not transferred to robustness, as the accuracy dropped to 0.33% under PGD, and single-step adversarial fine-tuning only increased the accuracy against FGSM, and the initial detector only identified 56.2% of attacked images, which means that the model cannot be called robust unless multi-step adversarial training is adopted or stronger defence strategies are introduced. Also, in some cases, network attention is on the annotation boxes, borders, and time stamps, pointing to partial shortcut learning, which can also hinder cross-dataset generalization [14]. The following limitations of the approaches are important to take into consideration when interpreting these findings: White-box attacks at the same perturbation size, same CBAM head for all models (can't distinguish the effect of the attention mechanism alone), a small test set of 600 images and a single public data set was used, and image-level split without patient identifier was used. With a 30.44 M parameter count and 32.59 ms per image, the model can also be used for offline analysis, but is not suitable for embedded applications.

6 Conclusion

The MaxViT-Tiny was proposed in this study to classify 8-class gastrointestinal endoscopic images based on a deduplicated Kvasir with a fixed stratified split with 3 different seeds. The mean accuracy was $93.72 \pm 1.06\%$ across the seeds, and was statistically significantly greater than the MobileNetV3 model in all seeds and greater

than the ConvNeXt-Tiny + GhostNet + DenseNet121 models in some seeds and statistically comparable to the ConvNeXt-Tiny + Swin-Tiny hybrid. Visually similar pairs of classes tended to be grouped in the residual errors, especially esophagitis and normal-z-line.

Results of the study showed that a high clean accuracy is not sufficient. Multi-step attacks were very successful against the model, and single-step attacks (adversarial fine-tuning) did not significantly reduce their attack rate: the pre-trained attack detector only succeeded in correctly identifying less than half of the attacked images, and the explanation maps revealed that the model used artefacts in the images. It is therefore important to be aware that this is a research prototype and not a clinically ready tool.

Future directions include (1) patient-level splits, cross-dataset testing and validation at multiple centres or in the prospective setting; (2) multi-step adversarial training and more powerful detection or certified defenses; (3) removing image overlays and borders and quantitatively assessing explanations; (4) reducing the model size for real-time and embedded usage; and (5) evaluating the backbones with and without the attention module, repeated over several seeds with paired significance tests, to quantify the contribution of CBAM.

Acknowledgements. Not applicable.

Declarations

- Funding No funding was received for this study.
- Conflict of interest/Competing interests The authors declare that they have no competing interests.
- Ethics approval and consent to participate Not applicable. This study used publicly available datasets and involved no new data collection or human participant recruitment.
- Consent for publication Not applicable.
- Data availability The datasets analyzed in this study are publicly available. The Kvasir dataset (version 1; 4,000 images, eight classes) is available from the Simula Research Laboratory at https://datasets.simula.no/kvasir/. GastroVision (8,000 images, 27 classes) is available from the Open Science Framework at https://osf.io/84e7f/. The copy used in this work was obtained from the Kaggle mirror at https://www.kaggle.com/datasets/orvile/gastrovision-gastrointestinal-disease-detection. No new data were collected for this study.
- Code availability The code used in this study is available from the corresponding author upon reasonable request.
- Author contribution Sadek Al Prince and Israt Fatema Shorna contributed equally to this work. Sadek Al Prince: conceptualization, methodology, software, formal analysis, investigation, and writing—original draft. Israt Fatema Shorna: methodology, validation, investigation, and writing—review and editing. Md Mehedi Hasan, Mohammad Ahanaf Akif, and Aziz Misher Mishu: validation, interpretation, and

writing—review and editing; Gazi Mehraj Sakib: writing—review and editing. All authors read and approved the final manuscript.

References

[1] Lewis, J.R., Pathan, S., Kumar, P. and Dias, C.C., 2024. AI in endoscopic gastrointestinal diagnosis: A systematic review of deep learning and machine learning techniques. IEEE Access, 12, pp.163764-163786.

[2] Rahi, P., Hussain, M.R., Alqahtani, S., Husain, M., Singh, A.P. and Singh, I., 2025. A Novel Random Forest-SMOTE Framework With Polynomial Feature Engineering for Early Detection of Gastrointestinal Disorders. IEEE Access, 13, pp.188577-188604.

[3] Shi, H., Wang, H., Ding, X. and Dang, Z., 2025. An in-depth study of personalized anesthesia management models in gastrointestinal endoscopy based on multimodal deep learning. IEEE Access, 13, pp.15410-15429.

[4] Simmonds, S., Cheng, L.K., Ruha, W.W., Taberner, A.J., Du, P. and Angeli-Gordon, T.R., 2023. Measurement and analysis of in vivo gastroduodenal slow wave patterns using anatomically-specific cradles and electrodes. IEEE Transactions on Biomedical Engineering, 71(4), pp.1289-1297.

[5] Ye, B., Shu, Z., Wang, B., Wang, S., Fu, Y., Zhang, L., Qi, B., Dwivedi, A.K. and Liu, S., 2024. Attention mechanism guided SE+ ResNet-H model for gastrointestinal endoscopy image classification. IEEE Transactions on Instrumentation and Measurement, 73, pp.1-13.

[6] Elmagzoub, M.A., Kaur, S., Gupta, S., Rajab, A., Rajab, K.D., Al Reshan, M.S., Alshahrani, H. and Shaikh, A., 2024. Improving endoscopic image analysis: Attention mechanism integration in grid search fine-tuned transfer learning model for multi-class gastrointestinal disease classification. IEEE Access, 12, pp.80345-80358.

[7] Mirza, O.M., Alsobhi, A., Hasanin, T., Ishak, M.K., Karim, F.K. and Mostafa, S.M., 2023. Computer-aided diagnosis for gastrointestinal cancer classification using hybrid rice optimization with deep learning. IEEE Access, 11, pp.76321-76329.

[8] Alruban, A., Alabdulkreem, E., Eltahir, M.M., Alharbi, A.R., Issaoui, I. and Sayed, A., 2023. Endoscopic image analysis for gastrointestinal tract disease diagnosis using a nature-inspired algorithm with a deep learning approach. IEEE Access, 11, pp.130022-130030.

[9] Sharmila, V. and Geetha, S., 2024. Gastrointestinal disease classification using hierarchical spatio pyramid transformer with pittree fusion and efficient-condconv swishnet. IEEE Access, 12, pp.113972-113987.

[10] Özbay, E., 2024. Gastrointestinal tract disease classification using residual-inception transformer with wireless capsule endoscopy images segmentation. IEEE Access, 12, pp.197988-197998.

[11] Ahmad, S., Kim, J.-S., Park, D.K. and Whangbo, T., 2023. Automated detection of gastric lesions in endoscopic images by leveraging attention-based YOLOv7. IEEE Access, 11, pp.87166-87177.

[12] Tran, T.H., Vu, D.H., Tran, M.H., Dao, V.H., Vu, H. and Nguyen, T.T., 2024. GIFCOS-DT: One stage detection of gastrointestinal tract lesions from endoscopic images with distance transform. IEEE Access, 12, pp.163698-163714.

[13] Garbaz, A., Oukdach, Y., Charfi, S., El Ansari, M., Koutti, L. and Salihoun, M., 2026. HDEFG-UFormer: A Hierarchical Depthwise-Expanded Feature Grouping Transformer-Based UNet for Gastrointestinal Disease Segmentation. IEEE Transactions on Systems, Man, and Cybernetics: Systems.

[14] Zhang, Y., Diao, Y.M., Yang, H. and Huang, Z.H., 2026. A Perturbation-Aware Local-to-Global Supervised Contrastive Framework for Robust Gastrointestinal Endoscopic Disease Classification. IEEE Access.

[15] Kanca, E., Gulsoy, T., Ayas, S. and Kablan, E.B., 2024, October. Generating robust adversarial images in medical image classification using GANs. In 2024 Innovations in Intelligent Systems and Applications Conference (ASYU) (pp. 1-5). IEEE.

[16] Manzari, O.N., Ahmadabadi, H., Kashiani, H., Shokouhi, S.B. and Ayatollahi, A., 2023. MedViT: a robust vision transformer for generalized medical image classification. Computers in Biology and Medicine, 157, p.106791.

[17] Obayya, M., Al-Wesabi, F.N., Maashi, M., Mohamed, A., Hamza, M.A., Drar, S., Yaseen, I. and Alsaid, M.I., 2023. Modified salp swarm algorithm with deep learning-based gastrointestinal tract disease classification on endoscopic images. IEEE Access, 11, pp.25959-25967.

[18] Alam, M.J., Rashid, R.B., Fattah, S.A. and Saquib, M., 2022. Rat-capsnet: A deep learning network utilizing attention and regional information for abnormality detection in wireless capsule endoscopy. IEEE Journal of Translational Engineering in Health and Medicine, 10, pp.1-8.

[19] Wang, W., Yang, X. and Tang, J., 2023. Vision transformer with hybrid shifted windows for gastrointestinal endoscopy image classification. IEEE Transactions on Circuits and Systems for Video Technology, 33(9), pp.4452-4461.

[20] Varam, D., Mitra, R., Mkadmi, M., Riyas, R.A., Abuhani, D.A., Dhou, S. and Alzaatreh, A., 2023. Wireless capsule endoscopy image classification: An explainable AI approach. IEEE Access, 11, pp.105262-105280.

[21] Agrusa, A.S., Kunkel, D.C. and Coleman, T.P., 2022. Robust regression and optimal transport methods to predict gastrointestinal disease etiology from high resolution EGG and symptom severity. IEEE Transactions on Biomedical Engineering, 69(11), pp.3313-3325.

[22] Kadian, S., Gopalakrishnan, S., Selvamani, V., Khan, S., Meyer, T., Thomas, R., Rana, M.M., Irazoqui, P.P., Verma, M.S. and Rahimi, R., 2023. Smart capsule for targeted detection of inflammation levels inside the GI tract. IEEE Transactions on Biomedical Engineering, 71(5), pp.1565-1576.

[23] Fontana, S., D'Agostino, S., Paffi, A., Marracino, P., Balucani, M., Ruocco, G., Aglioti, S.M., Apollonio, F. and Liberti, M., 2024. State of the art on advancements in wireless capsule endoscopy telemetry: a systematic approach. IEEE Open Journal of Antennas and Propagation, 5(5), pp.1282-1294.

[24] Li, Y., Zhang, T., Li, G., Zhang, H., Liu, P., Wang, C., Zhang, Z., Zhou, N., Dong, B. and Lou, C., 2024. Detection of exhaled methane in gastrointestinal disease population based on TDLAS. IEEE Sensors Journal, 24(19), pp.31095-31103.

[25] Dehghani, S., Mohsenipour, P., Kiya, F.P., Mohammadi, M., Farahani, R. and Beigzadeh, B., 2026. Curiosity-Driven Reinforcement Learning for Autonomous Navigation in Magnetically Controlled Capsule Endoscopy. IEEE Access.

[26] Qasaimeh, M., Ali, M. and Al-Haija, Q.A., 2026. GreenShield: A Lightweight and Robust Vision Transformer Framework in Retinal Disease Classification. Computer Modeling in Engineering & Sciences, 147(2), p.42.

[27] Neha, A.S., Chaturvedi, V. and Shafique, M., 2025, June. Hy-Deft: A Hybrid Defense Technique for Vision Transformers against Adversarial Attacks in Medical Imaging. In 2025 International Joint Conference on Neural Networks (IJCNN) (pp. 1-8). IEEE.

[28] Moligi, S., Dadi, R. and Atluri, V.V., 2025, November. MAViT: A Hybrid Adversarial Vision Transformer for Secure and Robust Threat Detection in IoT-based Medical Systems. In 2025 9th International Conference on Electronics, Communication and Aerospace Technology (ICECA) (pp. 516-523). IEEE.

[29] Pogorelov, K., Randel, K.R., Griwodz, C., Eskeland, S.L., de Lange, T., Johansen, D., Spampinato, C., Dang-Nguyen, D.-T., Lux, M., Schmidt, P.T., Riegler, M. and Halvorsen, P., 2017. Kvasir: A multi-class image dataset for computer-aided gastrointestinal disease detection. In Proceedings of the 8th ACM on Multimedia Systems Conference (MMSys), pp.164-169.

[30] Tu, Z., Talebi, H., Zhang, H., Yang, F., Milanfar, P., Bovik, A. and Li, Y., 2022. MaxViT: Multi-axis vision transformer. In Proceedings of the European Conference on Computer Vision (ECCV), pp.459-479.

[31] Woo, S., Park, J., Lee, J.-Y. and Kweon, I.S., 2018. CBAM: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), pp.3-19.

[32] Wightman, R., 2019. PyTorch Image Models (timm). GitHub repository, doi: 10.5281/zenodo.4414861. [Online]. Available: https://github.com/huggingface/pytorch-image-models

[33] Loshchilov, I. and Hutter, F., 2019. Decoupled weight decay regularization. In Proceedings of the International Conference on Learning Representations (ICLR).

[34] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. and Wojna, Z., 2016. Rethinking the Inception architecture for computer vision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.2818-2826.

[35] Goodfellow, I.J., Shlens, J. and Szegedy, C., 2015. Explaining and harnessing adversarial examples. In Proceedings of the International Conference on Learning Representations (ICLR).

[36] Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T. and Xie, S., 2022. A ConvNet for the 2020s. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp.11976-11986.

[37] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S. and Guo, B., 2021. Swin Transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp.10012-10022.

[38] Huang, G., Liu, Z., Van Der Maaten, L. and Weinberger, K.Q., 2017. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.4700-4708.

[39] Han, K., Wang, Y., Tian, Q., Guo, J., Xu, C. and Xu, C., 2020. GhostNet: More features from cheap operations. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp.1580-1589.

[40] Howard, A., Sandler, M., Chu, G., Chen, L.-C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., Le, Q.V. and Adam, H., 2019. Searching for MobileNetV3. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp.1314-1324.

[41] Cohen, J., 1960. A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20(1), pp.37-46.

[42] McNemar, Q., 1947. Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika, 12(2), pp.153-157.

[43] Madry, A., Makelov, A., Schmidt, L., Tsipras, D. and Vladu, A., 2018. Towards deep learning models resistant to adversarial attacks. In Proceedings of the International Conference on Learning Representations (ICLR).

[44] Dong, Y., Liao, F., Pang, T., Su, H., Zhu, J., Hu, X. and Li, J., 2018. Boosting adversarial attacks with momentum. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.9185-9193.

[45] Carlini, N. and Wagner, D., 2017. Towards evaluating the robustness of neural networks. In Proceedings of the IEEE Symposium on Security and Privacy (SP), pp.39-57.

[46] Liu, W., Wang, X., Owens, J. and Li, Y., 2020. Energy-based out-of-distribution detection. In Advances in Neural Information Processing Systems (NeurIPS), vol. 33, pp.21464-21475.

[47] Chattopadhay, A., Sarkar, A., Howlader, P. and Balasubramanian, V.N., 2018. Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks. In Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV), pp.839-847.

[48] Wang, H., Wang, Z., Du, M., Yang, F., Zhang, Z., Ding, S., Mardziel, P. and Hu, X., 2020. Score-CAM: Score-weighted visual explanations for convolutional neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp.24-25.

[49] Lundberg, S.M. and Lee, S.-I., 2017. A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems (NeurIPS), vol. 30, pp.4765-4774.

[50] Kokhlikyan, N., Miglani, V., Martin, M., Wang, E., Alsallakh, B., Reynolds, J., Melnikov, A., Kliushkina, N., Araya, C., Yan, S. and Reblitz-Richardson, O., 2020. Captum: A unified and generic model interpretability library for PyTorch. arXiv preprint arXiv:2009.07896.

[51] Jha, D., Sharma, V., Dasu, N., Tomar, N.K., Hicks, S., Bhuyan, M.K., Das, P.K., Riegler, M.A., Halvorsen, P., Bagci, U. and de Lange, T., 2023. GastroVision: A Multi-class Endoscopy Image Dataset for Computer-Aided Gastrointestinal Disease Detection. In Workshop on Machine Learning for Multimodal Healthcare Data (ML4MHD), ICML 2023. Dataset: https://osf.io/84e7f/; arXiv:2307.08140.

[52] Matthews, B.W., 1975. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochimica et Biophysica Acta (BBA) - Protein Structure, 405(2), pp.442-451.

[53] Guo, C., Pleiss, G., Sun, Y. and Weinberger, K.Q., 2017. On calibration of modern neural networks. In Proceedings of the International Conference on Machine

Learning (ICML), pp.1321-1330.