ADBA
Computers and Electronics in Medicine
Cutting-Edge Scientific Solutions

e-ISSN: 3023-8609
RESEARCH ARTICLE
Vol.3 / No.2 / 2026 / pp.162-173
https://doi.org/10.69882/adba.cem.2026077

Comparative Evaluation of Twenty Vision Transformer Variants for Six-Class Gastrointestinal Endoscopic Image Classification

Yousuf Islam [*,1], **Chaouki Aouiti** [α,2], **Asma Ladjeroud** [β,3] **and Meshkat Ahmad** [§,4]

[*]School of Microelectronics, University of Science and Technology of China, Hefei, China, [α]Department of Mathematics, University of Carthage, Bizerte, Tunisia, [β]Department of Mathematics, Blida 1 University, Blida, Algeria, [§]College of Biomedical Engineering, Fudan University, Shanghai, China.

ABSTRACT This study compares 20 vision transformer variants for six-class gastrointestinal endoscopic image classification using a labeled subset of HyperKvasir. The evaluated classes were cecum, polyps, pylorus, retroflex-rectum, retroflex-stomach, and z-line. Repeated row totals from six confusion matrices indicated an evaluation set of 768 images. The experiment was configured for five-fold cross-validation with a maximum of 50 epochs and a random seed of 42; however, the numerical fold-level file contained results only for Fold 1. Learning curves supplied qualitative information for selected runs in Folds 2–5 and complete graphical coverage for EViT-B0, EViT-B1, and EViT-B3. In the available numerical results, MaxViT-Tiny achieved the highest accuracy (99.09%), F1 score (99.10%), precision (99.12%), and recall (99.09%), whereas TinyViT-21M achieved the highest AUC (99.98%). MaxViT had the highest mean family accuracy, while TinyViT had the highest mean family AUC. Accuracy and F1 were strongly correlated, but the association between accuracy and AUC was weak. The learning curves showed rapid convergence with model-dependent validation fluctuations, and the confusion matrices located most errors in cecum–polyp and polyp–retroflex-rectum pairs. These findings support further evaluation of hybrid local–global transformer designs. Complete fold-level results, patient-level split information, calibration analysis, and external validation are required before drawing conclusions about cross-fold or clinical generalizability.

KEYWORDS
Gastrointestinal endoscopy
Hyperkvasir
Vision transformer
Image classification
Comparative evaluation

INTRODUCTION

Transformers changed sequence modeling by replacing recurrence with content-dependent attention (Vaswani *et al.* 2017). Vision Transformer (ViT) subsequently demonstrated that an image can be represented as a sequence of patches and processed by a largely unmodified transformer encoder (Dosovitskiy *et al.* 2021). Data-efficient training, hierarchical shifted windows, and hybrid convolution–attention blocks then broadened the design space (Touvron *et al.* 2021; Liu *et al.* 2021; Tu *et al.* 2022; Wu *et al.* 2022; Touvron *et al.* 2022; Mehta and Rastegari 2022; Cai *et al.* 2023). The resulting families differ not only in parameter scale, but also in locality bias, token interaction pattern, spatial hierarchy, distillation strategy, and computational objective. These design choices can alter the trade-off between hard-label accuracy, probability ranking, and deployability.

A credible comparison therefore requires more than reporting the numerically highest score. Accuracy and F1 summarize thresholded decisions, whereas the area under the receiver-operating-characteristic curve (AUC) measures ranking across thresholds (Fawcett 2006; Saito and Rehmsmeier 2015; Chicco and Jurman 2020). When classes are imbalanced, aggregate scores may conceal class-specific failure, and high discrimination does not guarantee calibration or transportability (Kelly *et al.* 2019; Roberts *et al.* 2021; Mongan *et al.* 2020; Liu *et al.* 2020; Wiens *et al.* 2019; Varoquaux 2018; Varma and Simon 2006; He and Garcia 2009; Chicco and Jurman 2020). Reproducible evaluation further requires leakage-resistant partitioning, fold-local preprocessing, transparent model selection, retained out-of-fold predictions, and uncertainty intervals (Roberts *et al.* 2021; Mongan *et al.* 2020; Varoquaux 2018; Varma and Simon 2006).

We evaluated 20 ViT-family models on six labeled categories from HyperKvasir: cecum, polyps, pylorus, retroflex-rectum, retroflex-stomach, and z-line (Borgli *et al.* 2020). The confusion matrices contained 768 observations from these categories. The

Manuscript received: 12 June 2026,
Revised: 8 July 2026,
Accepted: 15 July 2026.

[1]yislam15@mail.ustc.edu.cn (**Corresponding author**)
[2]chaouki.aouiti@fsb.rnu.tn
[3]ladjeroud_asma@univ-blida.dz
[4]meshkat22@m.fudan.edu.cn

experiment report specified 50 epochs, five folds, and a random seed of 42. All classifiers were labeled ViTHead, with no classical classifier or feature-selection stage reported. Numerical results were available only for Fold 1, whereas learning curves covered selected runs from the remaining folds and all five runs for EViT-B0, EViT-B1, and EViT-B3. We therefore used Fold 1 for quantitative model comparisons and the remaining curves only to describe training behavior. The experimental files did not document the complete subset-construction procedure, patient counts, or split assignments; consequently, 768 denotes the displayed evaluation set rather than the full HyperKvasir corpus.

The analysis addresses four questions: which model performs best in the available fold; whether the seven architecture families show consistent patterns across accuracy, F1, precision, recall, and AUC; whether accuracy and AUC produce similar model rankings; and which conclusions remain justified when numerical results are incomplete for Folds 2–5. All 20 models are reported, and numerical findings are distinguished from observations based only on learning curves.

RELATED WORKS

Canonical ViT uses global self-attention over patch tokens and benefits strongly from large-scale pretraining (Dosovitskiy *et al.* 2021). DeiT introduced data-efficient recipes and token-based distillation (Touvron *et al.* 2021), while DeiT III revisited augmentation, regularization, and optimization to strengthen ViT training without changing the central token-processing abstraction (Touvron *et al.* 2022). Swin Transformer instead constrains attention to shifted local windows organized in a hierarchy, supplying a stronger multiscale inductive bias (Liu *et al.* 2021). These strategies motivate direct comparison because they address the data efficiency and quadratic interaction cost of global attention in different ways.

MaxViT unifies convolution, blocked local attention, and dilated global attention within a hierarchical backbone (Tu *et al.* 2022). TinyViT reduces model scale through architecture constraints and fast pretraining distillation (Wu *et al.* 2022). MobileViTv2 replaces conventional multi-head attention with separable self-attention designed for linear token complexity (Mehta and Rastegari 2022), whereas EfficientViT uses lightweight multi-scale attention for efficient high-resolution processing (Cai *et al.* 2023). The benchmark therefore spans a meaningful continuum: global patch transformers, data-efficient training variants, windowed hierarchies, local–global hybrids, distilled compact transformers, and mobile-oriented attention.

CNN architectures remain important comparators because convolution introduces locality and translation-related inductive biases (He *et al.* 2016; Huang *et al.* 2017; Tan and Le 2019; Deng *et al.* 2009; LeCun *et al.* 2015). The present results therefore describe differences within the transformer models evaluated here; they do not establish superiority over CNN or hybrid pipelines. A broader comparison would require compute-matched CNN baselines together with parameter counts, latency, memory use, and energy measurements.

Kvasir established a multi-class resource for computer-aided gastrointestinal disease detection (Pogorelov *et al.* 2017), and HyperKvasir extended it with labeled and unlabeled images and videos (Borgli *et al.* 2020). The present benchmark uses five HyperKvasir anatomical or procedural categories—cecum, pylorus, retroflex-rectum, retroflex-stomach, and z-line—together with the polyp category. Kvasir-SEG is related to the polyp class but addresses localization through expert-verified masks rather than the six-class classification task considered here (Jha *et al.* 2020).

Endoscopic classification studies increasingly combine transfer learning, explanation, and self-supervision. One study paired ResNet152 with Grad-CAM on Kvasir images (Mukhtorov *et al.* 2023), while another used curriculum self-supervised learning to exploit unlabeled gastrointestinal endoscopy data (Guo *et al.* 2024). These studies underscore two requirements for the present transformer benchmark: errors must be localized to clinically meaningful classes, and high internal accuracy must be accompanied by explanation, calibration, and external validation before clinical claims are made.

High-stakes imaging literature has repeatedly shown that strong internal discrimination may fail under changes in acquisition source, prevalence, or institutional context (Litjens *et al.* 2017; Shen *et al.* 2017; Shamshad *et al.* 2023; Azad *et al.* 2024; Hatamizadeh *et al.* 2022; Tajbakhsh *et al.* 2016; Shin *et al.* 2016; Esteva *et al.* 2017; Topol 2019; Kelly *et al.* 2019; Roberts *et al.* 2021; Mongan *et al.* 2020). For gastrointestinal endoscopy, this implies splitting at the patient or examination level, avoiding near-duplicate video frames across partitions, reporting per-class behavior, preserving an external site when available, and separating discrimination from calibration. CLAIM and AI reporting extensions further emphasize transparent data provenance, reference standards, and analysis protocols (Mongan *et al.* 2020; Liu *et al.* 2020).

Cross-validation is especially vulnerable when model or hyperparameter selection is performed on the same folds used for performance estimation (Varoquaux 2018; Varma and Simon 2006). Augmentation and optimization also affect comparative conclusions and must be applied consistently and disclosed fully (Shorten and Khoshgoftaar 2019; Cubuk *et al.* 2020; Loshchilov and Hutter 2019; Kingma and Ba 2015; Lin *et al.* 2017; He and Garcia 2009). ROC-based summaries are useful but can appear optimistic under imbalance, making precision–recall analysis and confusion-matrix-derived metrics necessary complements (Fawcett 2006; Saito and Rehmsmeier 2015; Chicco and Jurman 2020). Explanation methods such as Grad-CAM, SHAP, and LIME may support failure analysis, but explanation plausibility is not a substitute for predictive validation (Selvaraju *et al.* 2017; Lundberg *et al.* 2020; Ribeiro *et al.* 2016). Transfer learning assumptions should likewise be documented because pretraining source and fine-tuning depth influence both performance and generalization (Pan and Yang 2010).

Among the requested authors, only two studies were retained because they are methodologically connected to medical-image classification. A kidney-image study compared deep-learning feature extractors combined with classical machine-learning classifiers (Alaca and Emin 2024), and a skin-cancer study used CNN features with genetic and particle-swarm optimization (Başaran and Çelik 2024). These studies motivate hybrid medical-image baselines, but neither provides a directly comparable gastrointestinal-endoscopy result.

MATERIALS AND METHODS

The available materials comprised an HTML experiment report, model-level and fold-level CSV files, 25 training and validation curve panels, six confusion matrices, six class-specific ROC plots, and five graphics labeled as K-fold summaries. The HTML report specified "Epoch: 50 | K-Fold: 5 | Seed: 42." The model summary included 20 rows and the fields Stage, ViT, Family, Variant, Algo, Classifier, Selected_Features, Mean_Acc, Std_Acc, Acc, F1, Precision, Recall, and AUC. Although the experiment was configured for five folds, every row in the fold-level file was labeled Fold = 1. Quantitative comparisons were therefore restricted to Fold 1. Curves from Folds 2–5 were used to examine convergence,

including the complete graphical series for EViT-B0, EViT-B1, and EViT-B3, but were not used to estimate missing metrics. Values were not inferred from unlabeled plot positions, and zero standard deviations based on a single record were not interpreted as evidence of reproducibility.

Dataset

The experiments used a six-class subset of the labeled HyperKvasir image collection, comprising cecum, polyps, pylorus, retroflex-rectum, retroflex-stomach, and z-line. HyperKvasir contains 110,079 gastrointestinal endoscopy images, including 10,662 labeled and 99,417 unlabeled images, and was released under the Creative Commons Attribution 4.0 license (Borgli *et al.* 2020). The experimental records supplied for this study identify 768 images in the displayed evaluation set, with class counts of 152, 154, 150, 58, 114, and 140, respectively. These counts describe the evaluated subset; the available files do not establish the number of training images, patient-level grouping, or the procedure used to select the six categories from the full dataset. Figure 1 presents representative endoscopic images from each category.

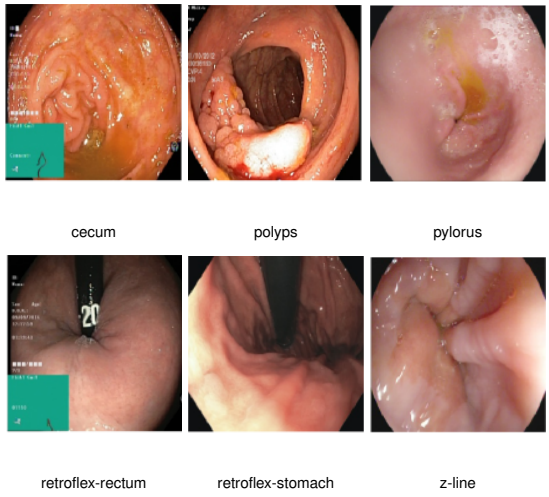

Figure 1 Representative images from the six HyperKvasir categories used in the benchmark. The panels are cropped from Figure 1 of (Borgli *et al.* 2020)

Table 1 Reconstructed class composition of the displayed 768-image evaluation set. Counts are repeated row totals from the supplied confusion matrices and do not establish the size of the full source corpus

Class	Images	Share	Clinical/visual role
cecum	152	19.79%	Anatomical landmark
polyps	154	20.05%	Pathological finding
pylorus	150	19.53%	Anatomical landmark
retroflex-rectum	58	7.55%	Procedural/anatomical view
retroflex-stomach	114	14.84%	Procedural/anatomical view
z-line	140	18.23%	Anatomical landmark
Total	**768**	**100.00%**	**Observed evaluation set**

Table 1 shows moderate class imbalance. Polyps is the largest displayed class (n = 154), whereas retroflex-rectum is the smallest (n = 58), yielding a largest-to-smallest ratio of 2.66. The remaining four classes contribute 114–152 images. This distribution makes overall accuracy sensitive to prevalence and supports reporting macro-averaged metrics and per-class recall. The six categories are part of the labeled HyperKvasir collection (Borgli *et al.* 2020). The counts in Table 1 describe the evaluation subset reconstructed from the supplied confusion matrices and should not be interpreted as the distribution of the complete HyperKvasir dataset.

Table 2 Experiment configuration and evidence status derived from the supplied report, CSV files, and graphical artifacts

Item	Recorded value	Evidentiary interpretation
Task	Six-class GI endoscopy classification	Confusion matrices identify cecum, polyps, pylorus, retroflex-rectum, retroflex-stomach, and z-line
Stage	VIT_ONLY	All 20 model rows
Architecture scope	20 variants; 7 families	Summary and fold-level CSV files
Epoch configuration	Maximum 50	HTML report; displayed curves terminate after 11–36 epochs
Configured K-fold	5	HTML report; configuration only
Numerical fold export	Fold 1 only	All 20 fold-level rows have Fold = 1
Graphical fold evidence	Folds 1–5	Complete curve sets for EViT-B0/B1/B3 and selected curves for other families; not a numerical metric matrix
Displayed confusion-matrix set	768 images per model	Class counts 152, 154, 150, 58, 114, and 140
Random seed	42	HTML report
Classifier	ViTHead	All model-level rows
Classical algorithm / selected features	None / not populated	Summary fields are empty
Std_Acc	0.0 for every model	Not a cross-fold stability estimate because one fold is present

Table 2 is central to the validity of the study. It distinguishes a five-fold configuration, graphical cross-fold learning evidence, and the one-fold numerical export available for model ranking. The new figures strengthen convergence and class-error analysis but do not make Std_Acc informative or permit confidence intervals and pairwise significance tests. The mismatch between the 50-epoch configuration and 11–36 displayed epochs is consistent with early stopping or another undocumented stopping rule; because that rule is not supplied, it remains an evidence-based inference rather than a reported hyperparameter. The benchmark includes ViT-Base, ViT-Small, and ViT-Tiny; DeiT3-Base and DeiT3-Small; Swin-Base, Swin-Small, and Swin-Tiny; TinyViT-21M, TinyViT-11M, and TinyViT-5M; MaxViT-Base, MaxViT-Small, and MaxViT-Tiny; EfficientViT variants EViT-B3, EViT-B1, and EViT-B0; and MobileViTv2 width multipliers 2.0, 1.0, and 0.5. The Variant field maps each model to Base, Small, or Tiny. Because the supplied artifacts do not expose pretraining source, input resolution, optimizer, scheduler, batch size, augmentation, or best-epoch rule, those details are not invented.

A recovered PyTorch checkpoint provided an independent structural check for the ViT-Base entry: 224 × 224 RGB input, 16 × 16 patch projection, 197 tokens including the class token, 768-dimensional embeddings, 12 encoder blocks, a 3072-dimensional MLP, a six-output head, and 85,803,270 parameters. The checkpoint confirms the task dimensionality and the configuration of one model, but does not establish that all families used identical preprocessing or pretraining. The supplied summary reports columns named Accuracy, F1, Precision, Recall, and AUC. For class c, precision is TPc/(TPc + FPc), recall is TPc/(TPc + FNc), and F1c is 2 × precisionc × recallc/(precisionc + recallc). Multiclass aggregation may be macro, micro, or prevalence-weighted, but the export does not identify the convention used for F1, precision, or recall. The same limitation applies to AUC: the file does not state whether the value is macro, micro, weighted, one-versus-rest, or one-versus-one. The manuscript therefore retains the generic metric labels and does not infer an averaging convention (Fawcett 2006; Saito and Rehmsmeier 2015; Chicco and Jurman 2020).

Models were ranked by accuracy, with F1 as the secondary

ordering field. Family means were computed as unweighted arithmetic means across the available variants. Overall ranges were expressed in percentage points. Pearson and Spearman correlations were calculated across the 20 model rows to describe metric concordance; these correlations are descriptive rather than confirmatory because model variants are related observations and arise from a single available fold. No p-values, confidence intervals, or multiple-comparison claims are used to declare a winning architecture. Learning curves were evaluated for convergence speed, train–validation separation, late-epoch oscillation, abrupt validation excursions, and fold-to-fold graphical consistency. Six confusion matrices were read in both count and row-normalized form to preserve the effect of class prevalence. Every matrix contains identical row totals, 152 cecum, 154 polyps, 150 pylorus, 58 retroflex-rectum, 114 retroflex-stomach, and 140 z-line images supporting direct error-pattern comparison on 768 evaluations. Six ROC figures were interpreted at the per-class level using the AUC values printed in each legend. The five K-fold summary graphics were treated as diagnostics of the export process because each contains a single Fold 1 bar despite its title.

RESULTS

Model and family performance

Across the 20 variants, mean accuracy was 98.15% (between-model SD 0.54 percentage points), mean F1 was 98.16%, mean precision was 98.20%, mean recall was 98.15%, and mean AUC was 99.851%. Accuracy ranged from 96.88% to 99.09%, a spread of 2.21 percentage points. AUC ranged from 99.59% to 99.98%, a narrower spread of 0.39 points. Table 3 shows that MaxViT-Tiny ranked first for every hard-label metric: 99.09% accuracy, 99.10% F1, 99.12% precision, and 99.09% recall. MaxViT-Base and MaxViT-Small ranked second and third by accuracy. TinyViT-21M, however, achieved the highest AUC (99.98%) while placing tenth by accuracy (98.31%). Conversely, MaxViT-Small ranked third by accuracy (98.70%) but had the lowest AUC (99.59%). These reversals demonstrate that a single metric cannot fully describe model behavior.

Table 3 Complete first-fold results for all 20 ViT-only models. Values are percentages; no completed five-fold mean or variance is implied

Rank	Model	Acc.	F1	Prec.	Recall	AUC
1	MaxViT-Tiny	99.09	99.10	99.12	99.09	99.89
2	MaxViT-Base	98.83	98.83	98.87	98.83	99.96
3	MaxViT-Small	98.70	98.71	98.75	98.70	99.59
4	Swin-Small	98.57	98.57	98.59	98.57	99.86
5	TinyViT-11M	98.44	98.44	98.44	98.44	99.97
6	TinyViT-5M	98.44	98.44	98.45	98.44	99.97
7	DeiT3-Base	98.31	98.32	98.40	98.31	99.87
8	Swin-Tiny	98.31	98.32	98.37	98.31	99.63
9	DeiT3-Small	98.31	98.31	98.35	98.31	99.81
10	TinyViT-21M	98.31	98.31	98.34	98.31	99.98
11	Swin-Base	98.18	98.18	98.25	98.18	99.90
12	EViT-B1	98.18	98.18	98.19	98.18	99.86
13	EViT-B0	98.18	98.18	98.21	98.18	99.79
14	MViTv2-1.0	98.18	98.18	98.20	98.18	99.86
15	EViT-B3	98.05	98.06	98.13	98.05	99.94
16	ViT-Tiny	97.79	97.80	97.83	97.79	99.84
17	MViTv2-2.0	97.79	97.79	97.85	97.79	99.92
18	ViT-Small	97.66	97.67	97.73	97.66	99.76
19	MViTv2-0.5	96.88	96.89	96.98	96.88	99.78
20	ViT-Base	96.88	96.88	96.96	96.88	99.84

Figure 2 shows the near overlap of accuracy and F1 within every model. This pattern is quantified by a Pearson correlation of

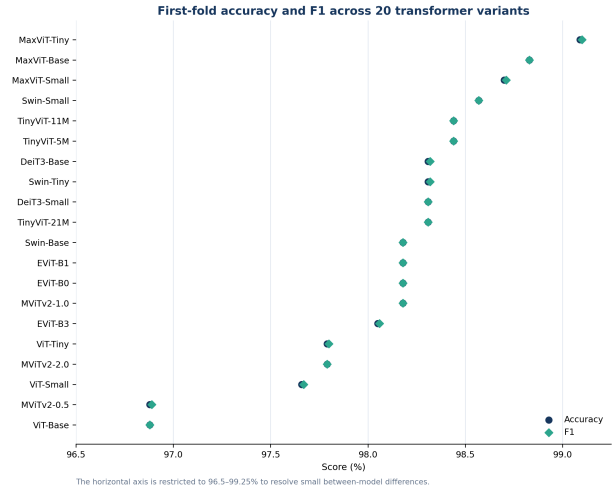

Figure 2 Ranked first-fold accuracy and F1 for all 20 transformer variants. The axis begins at 96.5% to make small differences visible; source: supplied model-level CSV

r = 0.99996, suggesting that the reported F1 does not materially reorder the systems relative to accuracy in Fold 1. The restricted axis makes differences visible but should not be mistaken for a large absolute separation: the entire accuracy spread is 2.21 percentage points. The concentration of MaxViT variants at the top nevertheless motivates confirmatory paired testing after all folds are recovered.

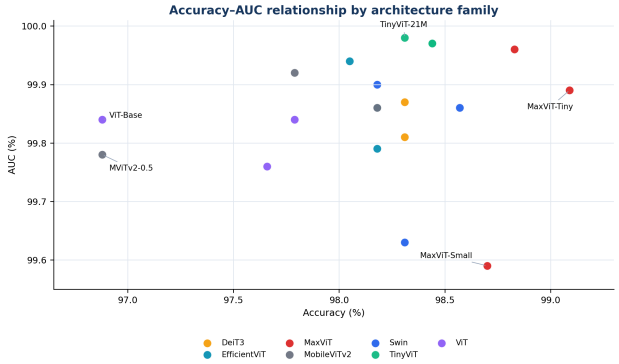

Figure 3 Relationship between first-fold accuracy and AUC across architecture families. Both axes are restricted to the observed high-performance region; source: supplied model-level CSV

Figure 3 indicates weak metric concordance. The descriptive Pearson association between accuracy and AUC was r = 0.127, and the Spearman rank association was ρ = 0.296. The most informative cases are TinyViT-21M and MaxViT-Small: the former led AUC without leading accuracy, whereas the latter combined top-three accuracy with the lowest AUC. This can arise when probability rankings are strong but the decision rule is suboptimal, or when high classification accuracy coexists with less consistent one-versus-rest ranking. Because the AUC aggregation convention and prediction probabilities are unavailable, the mechanism cannot be determined from the summary alone.

Table 4 ranks MaxViT first by mean accuracy (98.873%) and F1 (98.880%). TinyViT ranks second by mean accuracy (98.397%) but first by mean AUC (99.973%). Swin and DeiT3 occupy a narrow

Table 4 Unweighted family means across the variants available in Fold 1

Family	n	Acc. (%)	F1 (%)	Prec. (%)	Recall (%)	AUC (%)
MaxViT	3	98.873	98.880	98.913	98.873	99.813
TinyViT	3	98.397	98.397	98.410	98.397	99.973
Swin	3	98.353	98.357	98.403	98.353	99.797
DeiT3	2	98.310	98.315	98.375	98.310	99.840
EfficientViT	3	98.137	98.140	98.177	98.137	99.863
MobileViTv2	3	97.617	97.620	97.677	97.617	99.853
ViT	3	97.443	97.450	97.507	97.443	99.813

middle band around 98.31–98.35% mean accuracy. The canonical ViT family has the lowest mean accuracy (97.443%), followed by MobileViTv2 (97.617%). Family means are descriptive: families contain two or three non-exchangeable variants and therefore do not constitute replicated treatments.

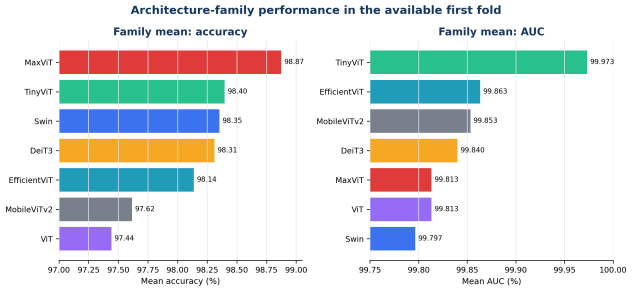

Figure 4 Family-level mean accuracy and AUC in the available first fold. Both panel axes are truncated to resolve the observed differences; source: supplied model-level CSV

Figure 4 summarizes the family-level trade-off. MaxViT benefits in hard-label classification, consistent with a design that combines local convolutional processing, blocked attention, and dilated global interaction (Tu *et al.* 2022). TinyViT's AUC advantage is compatible with highly discriminative but not necessarily optimally thresholded outputs (Wu *et al.* 2022). The figure does not establish causal architectural superiority because pretraining, parameter count, and hyperparameters are not controlled in the delivered metadata; rather, it identifies hypotheses for a compute-matched, completed-fold comparison.

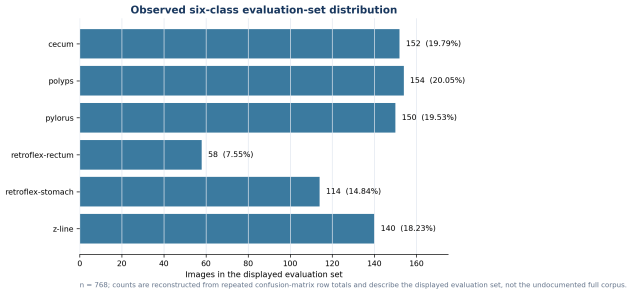

Figure 5 Observed class composition of the 768-image evaluation set reconstructed from repeated confusion-matrix row totals. The figure does not represent the undocumented full source corpus

Figure 5 presents the dataset evidence available at evaluation time. Cecum, polyps, and pylorus each account for approximately one fifth of the displayed set, whereas retroflex-rectum contributes only 7.55%. The 2.66-fold count difference between the largest and smallest classes is not extreme, but it is sufficient for micro-averaged or accuracy-based summaries to underweight rare-class errors. The figure therefore supports the use of class-wise recall and macro-averaged measures. It also enforces an important provenance boundary: n = 768 is the repeated confusion-matrix test count, not a verified full-dataset size, and no patient, procedure, center, or acquisition-device distribution was supplied.

Learning dynamics and class-wise diagnostics

The aggregate ranking does not reveal whether a model converged smoothly or passed through unstable validation states. Figures 6–12 therefore present 25 learning-curve panels, including complete five-fold graphical series for EViT-B0, EViT-B1, and EViT-B3 and representative fold curves for the other families. These figures are qualitative diagnostics because the numerical metrics for Folds 2–5 are unavailable and stopping epochs differ; no cross-model conclusion is based on visually reading an unreported numerical value.

Figure 6 Training and validation learning dynamics for the canonical ViT family: (a) ViT-Base, Fold 2; (b) ViT-Small, Fold 4; and (c) ViT-Tiny, Fold 4

Figure 6 shows rapid learning but variant-specific validation behavior. ViT-Base reaches high training accuracy while validation loss and accuracy oscillate near epochs 6 and 9, making checkpoint choice consequential. ViT-Small and ViT-Tiny maintain narrower train–validation gaps and steadier late-epoch loss, although validation accuracy still fluctuates. All three traces stop after approximately 11–13 epochs rather than the configured maximum of 50, indicating an undocumented stopping rule.

Figure 7 shows that a high final score can conceal transient validation instability. Swin-Base has the most pronounced event: validation loss spikes near epoch 10 while validation accuracy falls to roughly 96% before recovering. Swin-Small and Swin-Tiny also show intermittent validation-accuracy drops despite steadily high training accuracy. The recurrent recovery suggests that the models

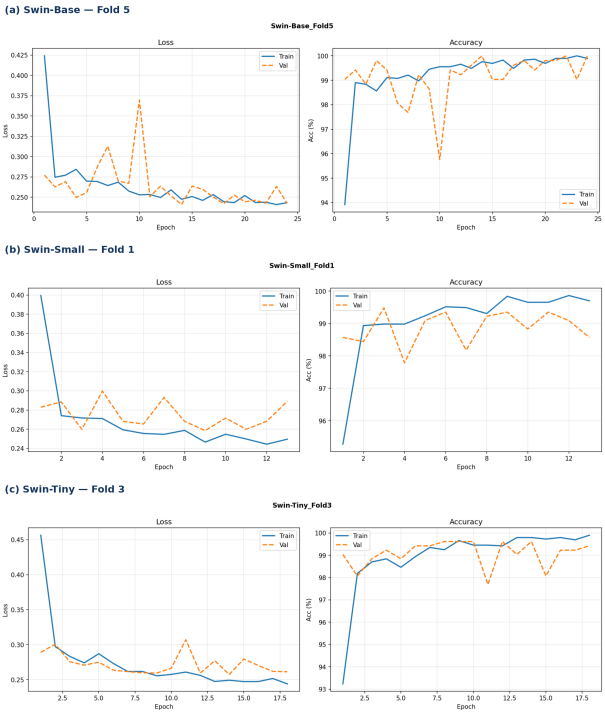

Figure 7 Training and validation learning dynamics for the Swin family: (a) Swin-Base, Fold 5; (b) Swin-Small, Fold 1; and (c) Swin-Tiny, Fold 3

remain trainable, but the oscillations make checkpoint-selection policy and patience settings methodologically consequential. Reporting only the terminal or best epoch would underrepresent this instability.

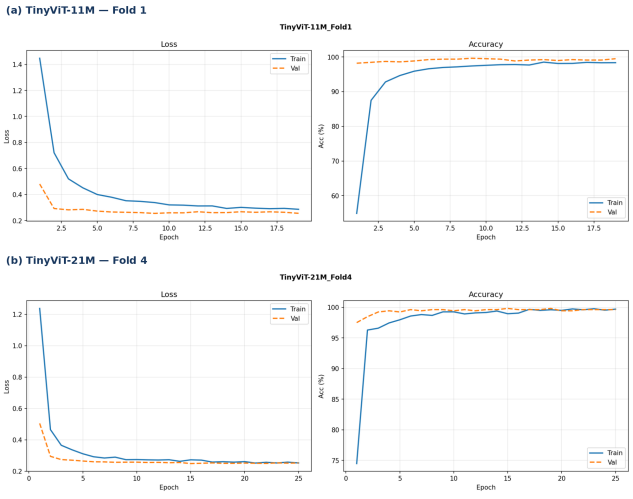

Figure 8 Training and validation learning dynamics for compact TinyViT models: (a) TinyViT-11M, Fold 1; and (b) TinyViT-21M, Fold 4

Figure 8 shows a steep reduction in training loss during the first few epochs for both TinyViT models, followed by a long plateau in which training and validation curves remain close. Validation accuracy is initially higher than training accuracy, a pattern compatible with stronger stochastic augmentation or regularization in the training path, although the supplied artifacts do not permit that mechanism to be confirmed. TinyViT-21M is particularly stable after early convergence, which is consistent with its leading AUC in Table 3 but does not by itself establish calibration or cross-fold superiority.

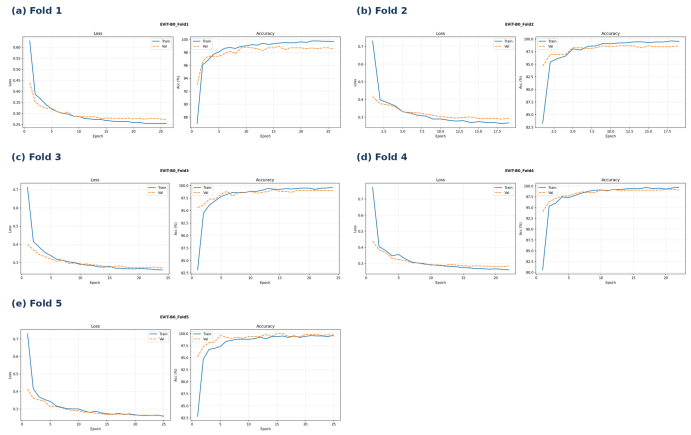

Figure 9 Training and validation learning dynamics for EViT-B0 across Folds 1–5

Figure 9 provides complete graphical fold coverage for EViT-B0. All five folds show a steep loss reduction during the first few epochs, followed by a lower-amplitude plateau in which training and validation losses remain close. Validation accuracy approaches the ceiling region without persistent collapse, although the exact stopping epoch varies across folds. This consistency supports optimization stability for EViT-B0 under the displayed runs; it does not yield a numerical cross-fold mean or variance because fold-specific metric rows and sample-level predictions are still absent.

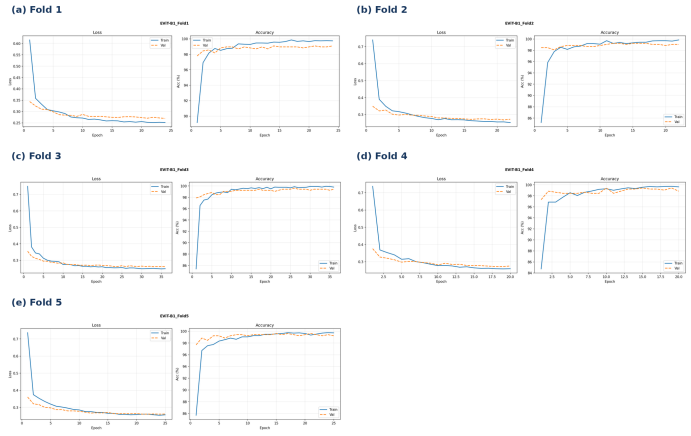

Figure 10 Training and validation learning dynamics for EViT-B1 across Folds 1–5

Figure 10 shows similarly rapid convergence for EViT-B1 across all five folds. Fold 3 continues for approximately 36 displayed epochs, whereas the remaining folds stop earlier, providing direct evidence that a common 50-epoch maximum was not always reached. Late-epoch validation loss generally remains above training loss by a small margin and validation accuracy stays close to training accuracy, suggesting limited but nonzero generalization gaps. The visual stability is compatible with the model's 98.18%. Fold 1 accuracy and 0.9986 aggregate AUC in Table 3, but it can not

establish that EViT-B1 is more reproducible than EViT-B0 without numerical fold outputs.

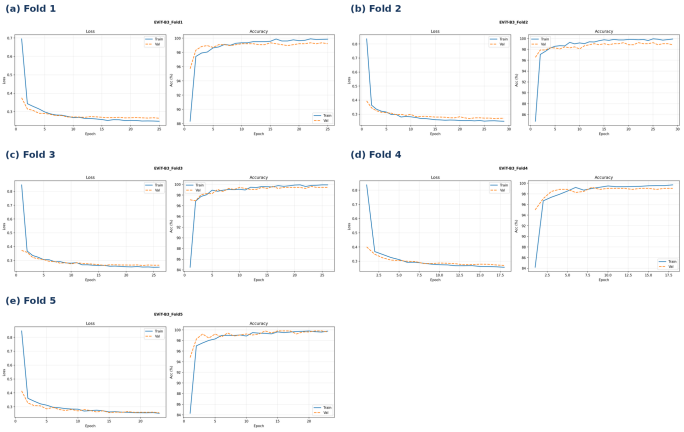

Figure 11 Training and validation learning dynamics for EViT-B3 across Folds 1–5

Figure 11 shows that EViT-B3 also converges rapidly in every fold despite higher initial training loss than validation loss. The train and validation trajectories become closely coupled after the early optimization phase, and validation accuracy remains in the high-performance region through the final displayed epochs. Fold-specific stopping ranges from roughly 18 to 29 epochs. The stability of ranking behavior is consistent with EViT-B3's highest EfficientViT AUC in Table 3, but the visible train–validation proximity does not substitute for patient-level leakage checks or external validation.

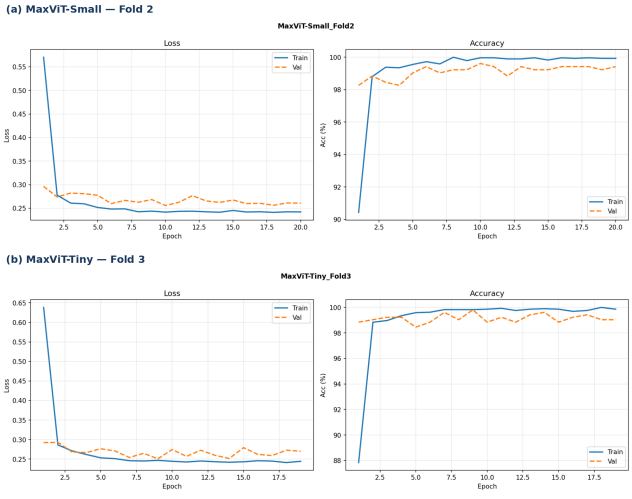

Figure 12 Training and validation learning dynamics for MaxViT models: (a) MaxViT-Small, Fold 2; and (b) MaxViT-Tiny, Fold 3

Figure 12 provides a convergence-level explanation for the strong MaxViT ranking. Both variants show a sharp early loss reduction, small late-epoch train–validation gaps, and validation accuracy that remains close to the ceiling after convergence. MaxViT-Tiny displays modest validation oscillation but no collapse, while MaxViT-Small maintains a stable validation-loss band around the mid-0.20s. The patterns are consistent with robust optimization in the displayed folds, but only complete fold-level predictions can establish whether this behavior yields lower generalization variance than competing families.

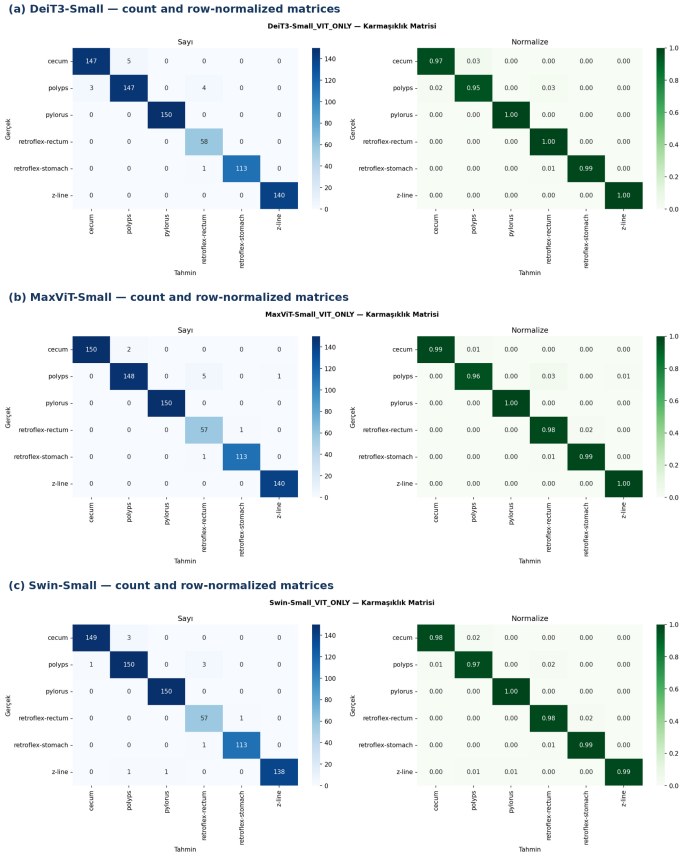

Figure 13 Count and row-normalized confusion matrices for (a) DeiT3-Small, (b) MaxViT-Small, and (c) Swin-Small on the 768-image evaluation set

Figure 13 confirms the six semantic classes and independently reproduces the corresponding Fold 1 accuracies: 755/768 (98.31%) for DeiT3-Small, 758/768 (98.70%) for MaxViT-Small, and 757/768 (98.57%) for Swin-Small. Pylorus is classified without error in all three matrices, whereas the principal residual confusions involve cecum versus polyps and polyps versus retroflex-rectum. MaxViT-Small reduces cecum-to-polyp errors to 2 but assigns five polyp images to retroflex-rectum and one to z-line. Swin-Small misclassifies two z-line images, while DeiT3-Small preserves perfect z-line recognition but has lower polyp recall. The repeated row totals also independently corroborate the class distribution visualized in Figure 5.

Figure 14 shows that EViT-B0 and EViT-B1 each classify 754 of 768 images correctly (98.18%), whereas EViT-B3 classifies 753 correctly (98.05%), exactly reproducing the corresponding Fold 1 accuracies in Table 3. EViT-B0 and EViT-B1 share the main cecum–polyp and polyp–retroflex-rectum confusions but differ in isolated pylorus, z-line, and retroflex-view errors. EViT-B3 raises polyp recall to 151/154 yet produces seven cecum-to-polyp errors and two retroflex-rectum-to-polyp errors. Thus, nearly identical aggregate accuracies conceal model-specific error topology, reinforcing the need for class-level reporting.

Figure 15 shows near-perfect one-versus-rest ranking for DeiT3-Base. Polyps has the lowest class-specific AUC (0.995), followed by retroflex-rectum (0.998) and cecum (0.999); pylorus, retroflex-stomach, and z-line reach 1.000 in the displayed evaluation. The

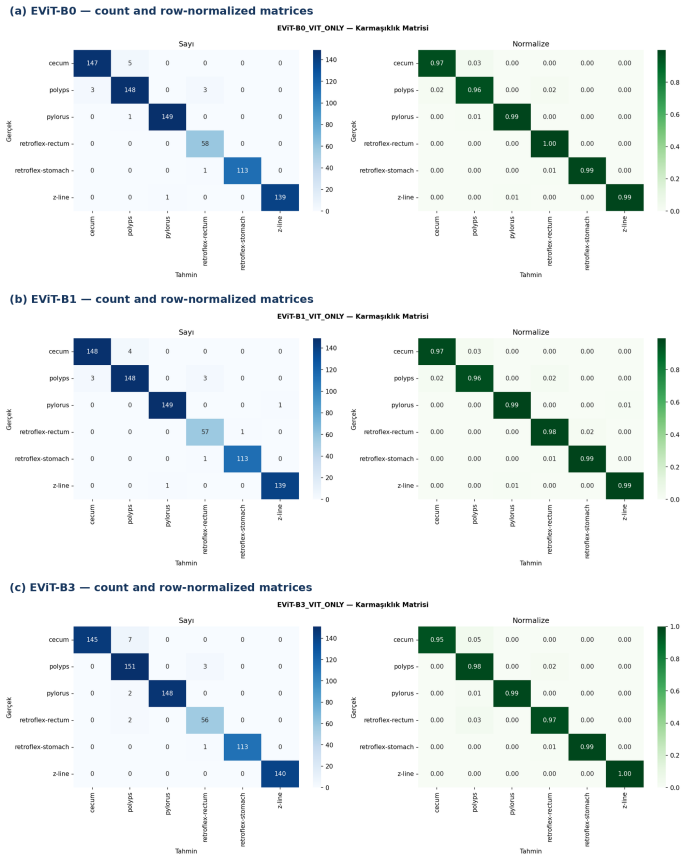

Figure 14 Count and row-normalized confusion matrices for (a) EViT-B0, (b) EViT-B1, and (c) EViT-B3 on the same 768-image evaluation set

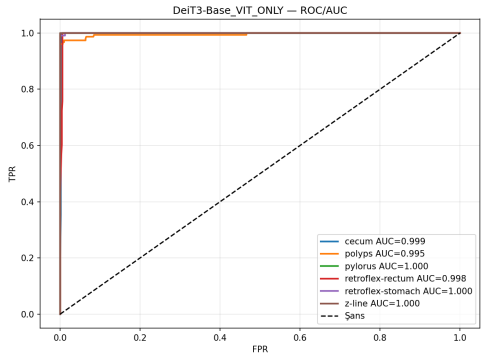

Figure 15 Class-specific ROC curves for DeiT3-Base. The printed AUC values range from 0.995 for polyps to 1.000 for pylorus, retroflex-stomach, and z-line

ordering is consistent with the confusion-matrix evidence that polyp-related decisions are the most difficult. Because these curves approach the upper-left corner, confidence intervals and an external test set are essential to distinguish genuine robustness from an optimistic internal split.

Figure 16 indicates uniformly high class-level discrimination for MaxViT-Base: every class lies between 0.999 and 1.000. This is compatible with the model-level AUC of 0.9996 in Table 3. The result does not contradict MaxViT-Base's lower accuracy than MaxViT-Tiny, because ROC AUC measures ranking across thresholds rather

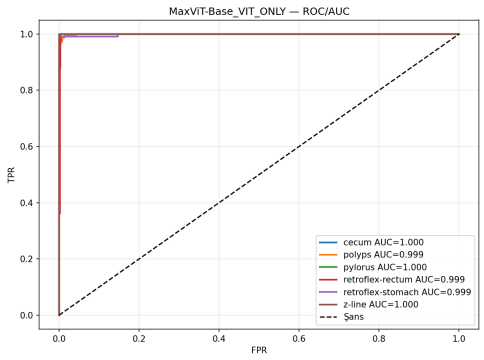

Figure 16 Class-specific ROC curves for MaxViT-Base. Cecum, pylorus, and z-line reach an AUC of 1.000; the remaining classes are reported as 0.999

than the correctness of the final multiclass decision. Calibration curves and decision-threshold documentation are required to explain the residual difference.

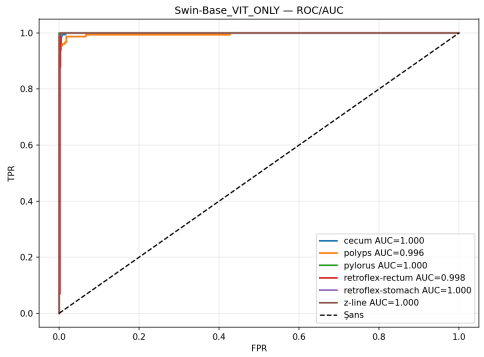

Figure 17 Class-specific ROC curves for Swin-Base. Polyps has the lowest displayed AUC (0.996), whereas cecum, pylorus, retroflex-stomach, and z-line reach 1.000

Figure 17 again identifies polyps as the comparatively difficult class, with AUC 0.996, while retroflex-rectum reaches 0.998 and four classes reach 1.000. The pattern parallels the error topology in Figure 13, where polyp images account for several off-diagonal counts. The ROC curves remain threshold-independent summaries; precision–recall curves would be an important complement because the class prevalences are unequal and the rare retroflex-rectum class contains only 58 displayed examples (Saito and Rehmsmeier 2015).

Figure 18 identifies the clearest EfficientViT ranking weakness: the polyp AUC of 0.991 is lower than the values for the remaining classes, while retroflex-rectum reaches 0.998, cecum 0.999, and pylorus, retroflex-stomach, and z-line 1.000. This class ordering is compatible with the cecum–polyp and polyp–retroflex-rectum errors in Figure 14. The very high values remain internal discrimination estimates and should be complemented by prevalence-sensitive precision–recall curves and calibration analysis.

Figure 19 shows a modest improvement over EViT-B0 in polyp ranking, from 0.991 to 0.994, while retaining 0.998 for retroflex-rectum and 1.000 for cecum, pylorus, retroflex-stomach, and z-line. This change helps explain why EViT-B1 has a higher aggregate AUC than EViT-B0 despite identical Fold 1 accuracy. It also illustrates why threshold-independent ranking and hard-label cor-

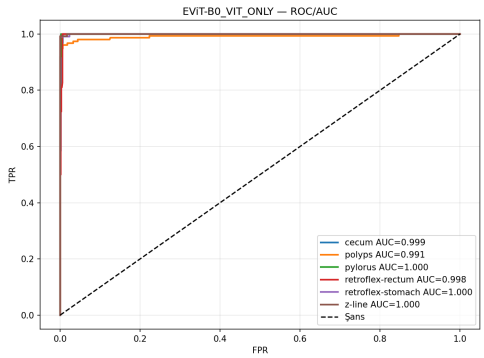

Figure 18 Class-specific ROC curves for EViT-B0. Polyps has the lowest displayed AUC (0.991); cecum reaches 0.999 and three classes reach 1.000

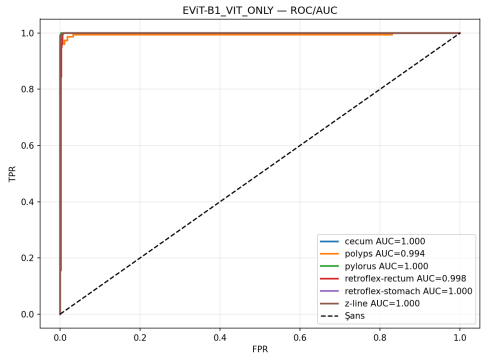

Figure 19 Class-specific ROC curves for EViT-B1. Polyps has an AUC of 0.994, retroflex-rectum 0.998, and the remaining four classes 1.000

rectness should be reported together rather than treated as interchangeable outcomes.

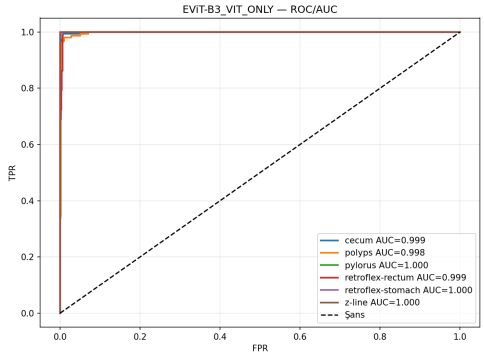

Figure 20 Class-specific ROC curves for EViT-B3. Polyps reaches 0.998, cecum and retroflex-rectum 0.999, and three classes 1.000

Figure 20 presents the strongest class-wise ranking profile among the three EfficientViT variants: polyps reaches 0.998, cecum and retroflex-rectum reach 0.999, and pylorus, retroflex-stomach, and z-line reach 1.000. This is consistent with EViT-B3's aggregate AUC of 0.9994, even though its hard-label accuracy is slightly lower than EViT-B0 and EViT-B1. The result is a concrete example of metric discordance caused by thresholded decisions rather than weak probability ordering.

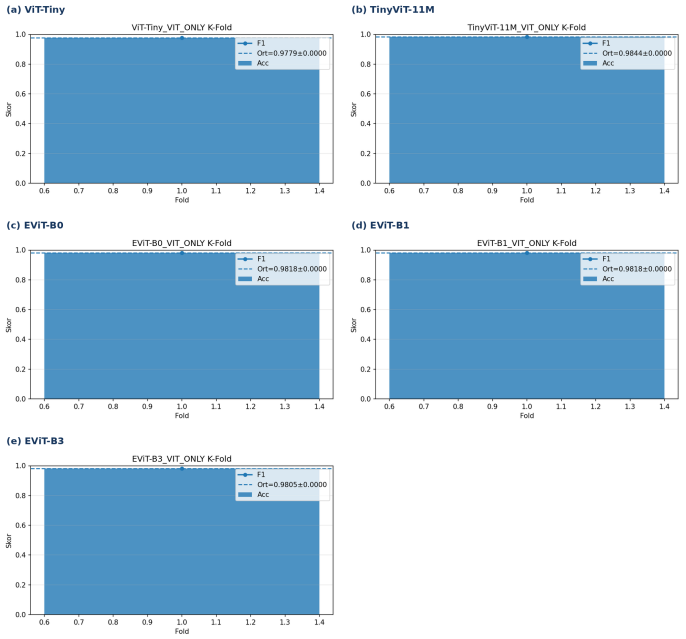

Figure 21 Supplied K-fold summary graphics for (a) ViT-Tiny, (b) TinyViT-11M, (c) EViT-B0, (d) EViT-B1, and (e) EViT-B3. Each graphic contains only Fold 1

Figure 21 clarifies the zero standard deviations in the summary output. Although the five graphics are labeled "K-Fold," each contains a single bar for Fold 1. The reported values are 0.9779, 0.9844, 0.9818, 0.9818, and 0.9805, each accompanied by ± 0.0000. A standard deviation based on one observation cannot describe variation across folds. These plots are therefore interpreted as evidence of an incomplete export, not as cross-validation summaries.

Every ranked value in Tables 3–4 and Figures 2–4 derives from Fold 1. Figure 5 reconstructs only the displayed evaluation-set composition. Figures 6–12 permit qualitative convergence assessment and provide complete graphical five-fold series for the three EfficientViT variants, while Figures 13–20 provide class-level Fold 1 evidence and Figure 21 documents the single-fold summary-export defect. None of these graphics supplies a complete 20-model-by-five-fold numerical matrix. The summary column Mean_Acc equals the Fold 1 accuracy rounded to four decimals, and Std_Acc equals 0.0 for all models. Consequently, no confidence interval, critical-difference diagram, corrected resampled test, or claim of statistically significant five-fold superiority is warranted.

DISCUSSION

Interpretation and model selection

MaxViT performed consistently well in the available fold. MaxViT-Tiny ranked first, and the three MaxViT variants occupied the top three positions for accuracy. This ordering suggests that the combination of local and global operations was well suited to the present data, although differences in parameter count and optimization settings could not be examined.

AUC values were concentrated near the upper limit, whereas accuracy produced a wider separation among models. TinyViT-21M had the highest AUC but ranked tenth in accuracy, and MaxViT-Small showed the reverse pattern. Model selection therefore depends on whether the intended use emphasizes final class assignments or probability ranking (Fawcett 2006; Saito and Rehmsmeier

2015). Calibration remains a separate question.

The canonical ViT variants generally ranked below the hierarchical and efficiency-oriented models in Fold 1. ViT-Base shared the lowest accuracy despite an AUC of 99.84%, and ViT-Small and ViT-Tiny were also below the overall mean. This result is consistent with reports that locality, hierarchy, data-efficient training, and distillation can improve practical vision models (Touvron *et al.* 2021; Liu *et al.* 2021; Tu *et al.* 2022; Wu *et al.* 2022; Touvron *et al.* 2022; Mehta and Rastegari 2022; Cai *et al.* 2023). Confirmation requires identical preprocessing and complete numerical results for all folds.

The learning curves provide information that is not captured by the final metric rows. MaxViT, TinyViT, and EfficientViT examples were comparatively stable after convergence, whereas several ViT and Swin runs showed noticeable validation fluctuations. The confusion matrices also showed that errors were not evenly distributed across classes: pylorus was classified consistently well, while most remaining errors involved polyps and adjacent anatomical views. These patterns support patient-level validation and class-specific review rather than reliance on a single aggregate score (Pogorelov *et al.* 2017; Borgli *et al.* 2020; Jha *et al.* 2020; Mukhtorov *et al.* 2023; Guo *et al.* 2024).

On the basis of Fold 1, MaxViT-Tiny is the strongest candidate for further evaluation, but the available evidence is insufficient to designate a definitive model. A confirmatory analysis should specify a primary outcome in advance, use the same folds and preprocessing for every architecture, and estimate pairwise uncertainty from predictions on the same samples. Calibration and computational cost should be considered alongside discrimination.

TinyViT, EfficientViT, and MobileViTv2 were designed for compact or latency-sensitive applications (Wu *et al.* 2022; Mehta and Rastegari 2022; Cai *et al.* 2023). The available records do not include parameter counts, FLOP estimates, memory use, or latency for the trained models. Efficiency should therefore be measured directly rather than inferred from architecture names.

Reproducibility and confirmatory analysis

A reproducible follow-up study would retain image and examination identifiers, patient-level fold assignments, acquisition information, preprocessing and augmentation code, model initialization details, training hyperparameters, random seeds, and software and hardware versions. For each model and fold, the true label, predicted class, class probabilities, and sample identifier should be stored. These records would support pooled out-of-fold analyses, calibration assessment, and patient-clustered uncertainty estimates.

Where repeated images arise from the same subject or object, splitting and bootstrap resampling must occur at that higher level. Hyperparameter search requires nested validation or a locked tuning split to avoid optimistic bias (Varoquaux 2018; Varma and Simon 2006). After one primary outcome is fixed, paired bootstrap confidence intervals or a suitable repeated-measures procedure can compare models; multiplicity control should be applied to a limited set of prespecified contrasts. External validation should remain untouched until the final model and thresholding procedure are frozen.

Two retained medical-image studies suggest a limited confirmatory extension: deep transformer features may be compared with classical classifiers under identical folds, following a kidney-image hybrid design and an optimized skin-image classification design (Alaca and Emin 2024; Başaran and Çelik 2024). These citations support a methodological comparison only and do not provide performance evidence for the present dataset.

LIMITATIONS

The most important limitation is the incomplete numerical fold export. Although the HTML report specifies five-fold cross-validation and the EViT-B0/B1/B3 curves document all five training runs graphically, the metric CSV contains only Fold 1; thus, generalization variability and statistical superiority cannot be estimated. Second, the evaluation categories are drawn from HyperKvasir, but the subset-selection procedure, patient and examination counts, and split granularity remain undocumented. The displayed matrices establish only a 768-image evaluation set with unequal class counts. Third, the multiclass averaging conventions for F1, precision, recall, and AUC are not documented. Fourth, training hyperparameters, stopping logic, pretraining source, and compute measurements are absent. Fifth, the family analysis averages heterogeneous model scales and remains exploratory. The high AUC values and nearly diagonal confusion matrices may reflect strong class separability. They also make it important to exclude near-duplicate video frames across splits and to verify performance at the patient level and at an external site. The available materials did not include sample-level probabilities, calibration measures, confidence intervals, or a documented error-review set. The results should therefore be regarded as an initial comparative analysis rather than evidence of clinical readiness.

CONCLUSION

Among the 20 models evaluated in Fold 1, MaxViT-Tiny achieved the highest accuracy, F1 score, precision, and recall, while TinyViT-21M achieved the highest AUC. At the family level, MaxViT had the highest mean accuracy and TinyViT the highest mean AUC. The weak association between accuracy and AUC indicates that class assignment and probability ranking describe different aspects of performance. Learning curves and class-level error analyses provided additional information about convergence and residual misclassification. These results support further evaluation of MaxViT-Tiny and the TinyViT variants, but they do not establish performance across all five folds or at an external site. Numerical results and sample-level predictions for Folds 2–5, together with patient-level split information, training details, calibration analysis, and computational measurements, are required for a confirmatory comparison.

Availability of data and material

The analysis used aggregate model results, Fold 1 numerical records, 25 learning-curve panels, six confusion matrices, six class-specific ROC plots, and five single-fold summary graphics. The full image corpus, patient or examination identifiers, sample-level predictions, and numerical results for Folds 2–5 were not available.

Funding

This research received no external funding.

Conflicts of interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Ethical standard

The authors have no relevant financial or non-financial interests to disclose.

LITERATURE CITED

Alaca, Y. and B. Emin, 2024 Performance evaluation of hybrid approaches combining deep learning models and machine learning methods for medical kidney image classification. 9th Azerbaijan Congress on Life, Engineering, Mathematical, and Applied Sciences p. 3007.

Azad, R., A. Kazerouni, M. Heidari, E. K. Aghdam, A. Molaei, *et al.*, 2024 Advances in medical image analysis with vision transformers: A comprehensive review. Medical Image Analysis **91**: 103000.

Başaran, E. and Y. Çelik, 2024 Skin cancer diagnosis using cnn features with genetic algorithm and particle swarm optimization methods. Transactions of the Institute of Measurement and Control **46**: 2733–2744.

Borgli, H., V. Thambawita, P. H. Smedsrud, S. Hicks, D. Jha, *et al.*, 2020 Hyperkvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy. Scientific Data **7**: 283.

Cai, H., C. Li, M. Hu, C. Gan, and S. Han, 2023 Efficientvit: Lightweight multi-scale attention for high-resolution dense prediction. Proceedings of the IEEE/CVF International Conference on Computer Vision pp. 17302–17313.

Chicco, D. and G. Jurman, 2020 The advantages of the matthews correlation coefficient over f1 score and accuracy in binary classification evaluation. BMC Genomics **21**: 6.

Cubuk, E. D., B. Zoph, J. Shlens, and Q. V. Le, 2020 Randaugment: Practical automated data augmentation with a reduced search space. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops pp. 3008–3017.

Deng, J., W. Dong, R. Socher, L.-J. Li, K. Li, *et al.*, 2009 Imagenet: A large-scale hierarchical image database. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition pp. 248–255.

Dosovitskiy, A., L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, *et al.*, 2021 An image is worth 16x16 words: Transformers for image recognition at scale. International Conference on Learning Representations .

Esteva, A., B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, *et al.*, 2017 Dermatologist-level classification of skin cancer with deep neural networks. Nature **542**: 115–118.

Fawcett, T., 2006 An introduction to roc analysis. Pattern Recognition Letters **27**: 861–874.

Guo, H., S. A. Somayajula, R. Hosseini, *et al.*, 2024 Improving image classification of gastrointestinal endoscopy using curriculum self-supervised learning. Scientific Reports **14**: 3641.

Hatamizadeh, A., Y. Tang, V. Nath, D. Yang, A. Myronenko, *et al.*, 2022 Unetr: Transformers for 3d medical image segmentation. IEEE/CVF Winter Conference on Applications of Computer Vision pp. 574–584.

He, H. and E. A. Garcia, 2009 Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering **21**: 1263–1284.

He, K., X. Zhang, S. Ren, and J. Sun, 2016 Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition pp. 770–778.

Huang, G., Z. Liu, L. van der Maaten, and K. Q. Weinberger, 2017 Densely connected convolutional networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition pp. 4700–4708.

Jha, D., P. H. Smedsrud, M. A. Riegler, D. Johansen, T. De Lange, *et al.*, 2020 Kvasir-seg: A segmented polyp dataset. International Conference on Multimedia Modeling pp. 451–462.

Kelly, C. J., A. Karthikesalingam, M. Suleyman, G. Corrado, and D. King, 2019 Key challenges for delivering clinical impact with artificial intelligence. BMC Medicine **17**: 195.

Kingma, D. P. and J. Ba, 2015 Adam: A method for stochastic optimization. International Conference on Learning Representations .

LeCun, Y., Y. Bengio, and G. Hinton, 2015 Deep learning. Nature **521**: 436–444.

Lin, T.-Y., P. Goyal, R. Girshick, K. He, and P. Dollár, 2017 Focal loss for dense object detection. Proceedings of the IEEE International Conference on Computer Vision pp. 2980–2988.

Litjens, G., T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, *et al.*, 2017 A survey on deep learning in medical image analysis. Medical Image Analysis **42**: 60–88.

Liu, X., S. C. Rivera, D. Moher, M. J. Calvert, and A. K. Denniston, 2020 Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: The consort-ai extension. Nature Medicine **26**: 1364–1374.

Liu, Z., Y. Lin, Y. Cao, H. Hu, Y. Wei, *et al.*, 2021 Swin transformer: Hierarchical vision transformer using shifted windows. Proceedings of the IEEE/CVF International Conference on Computer Vision pp. 10012–10022.

Loshchilov, I. and F. Hutter, 2019 Decoupled weight decay regularization. International Conference on Learning Representations .

Lundberg, S. M., G. Erion, H. Chen, A. DeGrave, S. M. Pruthi, *et al.*, 2020 From local explanations to global understanding with explainable ai for trees. Nature Machine Intelligence **2**: 56–67.

Mehta, S. and M. Rastegari, 2022 Separable self-attention for mobile vision transformers. arXiv preprint arXiv:2206.02680 .

Mongan, J., L. Moy, and J. Kahn, Charles E., 2020 Checklist for artificial intelligence in medical imaging (claim): A guide for authors and reviewers. Radiology: Artificial Intelligence **2**: e200029.

Mukhtorov, D., M. Rakhmonova, S. Muksimova, and Y.-I. Cho, 2023 Endoscopic image classification based on explainable deep learning. Sensors **23**: 3176.

Pan, S. J. and Q. Yang, 2010 A survey on transfer learning. IEEE Transactions on Knowledge and Data Engineering **22**: 1345–1359.

Pogorelov, K., K. R. Randel, C. Griwodz, S. L. Eskeland, T. de Lange, *et al.*, 2017 Kvasir: A multi-class image dataset for computer aided gastrointestinal disease detection. Proceedings of the 8th ACM Multimedia Systems Conference pp. 164–169.

Ribeiro, M. T., S. Singh, and C. Guestrin, 2016 Why should i trust you?: Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining pp. 1135–1144.

Roberts, M., D. Driggs, M. Thorpe, M. van der Schaar, C.-B. Schonlieb, *et al.*, 2021 Common pitfalls and recommendations for using machine learning to detect and prognosticate for covid-19 using chest radiographs and ct scans. Nature Machine Intelligence **3**: 199–217.

Ronneberger, O., P. Fischer, and T. Brox, 2015 U-net: Convolutional networks for biomedical image segmentation. Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015 **9351**: 234–241.

Saito, T. and M. Rehmsmeier, 2015 The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. PLOS ONE **10**: e0118432.

Selvaraju, R. R., M. Cogswell, A. Das, R. Vedantam, D. Parikh, *et al.*, 2017 Grad-cam: Visual explanations from deep networks via gradient-based localization. Proceedings of the IEEE International Conference on Computer Vision pp. 618–626.

Shamshad, F., S. Khan, S. W. Zamir, M. H. Khan, M. Hayat, *et al.*, 2023 Transformers in medical imaging: A survey. Medical Image Analysis **88**: 102802.
Shen, D., G. Wu, and H.-I. Suk, 2017 Deep learning in medical image analysis. Annual Review of Biomedical Engineering **19**: 221–248.
Shin, H.-C., H. R. Roth, M. Gao, L. Lu, Z. Xu, *et al.*, 2016 Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning. IEEE Transactions on Medical Imaging **35**: 1285–1298.
Shorten, C. and T. M. Khoshgoftaar, 2019 A survey on image data augmentation for deep learning. Journal of Big Data **6**: 60.
Tajbakhsh, N., J. Y. Shin, S. R. Gurudu, R. T. Hurst, C. B. Kendall, *et al.*, 2016 Convolutional neural networks for medical image analysis: Full training or fine tuning? IEEE Transactions on Medical Imaging **35**: 1299–1312.
Tan, M. and Q. V. Le, 2019 Efficientnet: Rethinking model scaling for convolutional neural networks. Proceedings of Machine Learning Research **97**: 6105–6114.
Topol, E. J., 2019 High-performance medicine: The convergence of human and artificial intelligence. Nature Medicine **25**: 44–56.
Touvron, H., M. Cord, M. Douze, F. Massa, A. Sablayrolles, *et al.*, 2021 Training data-efficient image transformers and distillation through attention. Proceedings of Machine Learning Research **139**: 10347–10357.
Touvron, H., M. Cord, and H. Jégou, 2022 Deit iii: Revenge of the vit. European Conference on Computer Vision pp. 516–533.
Tu, Z., H. Talebi, H. Zhang, F. Yang, M. Peyfe, *et al.*, 2022 Maxvit: Multi-axis vision transformer. European Conference on Computer Vision pp. 459–477.
Varma, S. and R. Simon, 2006 Bias in error estimation when using cross-validation for model selection. BMC Bioinformatics **7**: 91.
Varoquaux, G., 2018 Cross-validation failure: Small sample sizes lead to large error bars. NeuroImage **180**: 68–77.
Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, *et al.*, 2017 Attention is all you need. Advances in Neural Information Processing Systems **30**: 5998–6008.
Wiens, J., S. Saria, M. Sendak, M. Ghassemi, V. Liu, *et al.*, 2019 Do no harm: A roadmap for responsible machine learning for health care. Nature Medicine **25**: 1337–1340.
Wu, K., J. Zhang, H. Peng, M. Liu, B. Xiao, *et al.*, 2022 Tinyvit: Fast pretraining distillation for small vision transformers. European Conference on Computer Vision pp. 68–85.
Zech, J. R., M. A. Badgeley, M. Liu, A. B. Costa, J. J. Titano, *et al.*, 2018 Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs. PLOS Medicine **15**: e1002683.

How to cite this article: Islam, Y., Aouiti, C., Ladjeroud, A., and Ahmad, M. Comparative Evaluation of Twenty Vision Transformer Variants for Six-Class Gastrointestinal Endoscopic Image Classification. *Computers and Electronics in Medicine*, 3(2), 162–173, 2026.