Adaptive Inference for Medical Vision Transformers: Token Reduction or Early Exit?

Ji Young Byun * [1] JBYUN13@JHU.EDU
HyunSeo Lee *[1] HLEE267@JHU.EDU
Jordan Shuff [1,3,4,5] JSHUFF1@JHU.EDU
Rengaraj Venkatesh [6] VENKATESH@ARAVIND.ORG
Nakul S. Shekhawat† [4] NSHEKHA1@JHMI.EDU
Kunal S. Parikh† [1,3,4,5] KSP@JHU.EDU
Rama Chellappa† [1,2] RCHELLA4@JHU.EDU

[1] *Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD, USA*
[2] *Department of Electrical and Computer Engineering, Johns Hopkins University, Baltimore, MD, USA*
[3] *Glaucoma Center of Excellence and Center for Nanomedicine, Wilmer Eye Institute, Johns Hopkins University School of Medicine, Baltimore, MD, USA*
[4] *Wilmer Eye Institute, Johns Hopkins University School of Medicine, Baltimore, MD, US*
[5] *Center for Bioengineering Innovation & Design, Johns Hopkins University, Baltimore, MD, USA*
[6] *Aravind Eye Hospital, Pondicherry, India*

Editors: Under Review for MIDL 2026

Abstract

Vision Transformers (ViTs) have demonstrated exceptional performance in medical image analysis, yet their computational demands hinder clinical deployment, particularly in time-sensitive applications. Medical imaging requires sample-adaptive optimization due to dataset heterogeneity across modalities and sample complexity; uniform strategies do not well balance efficiency and accuracy. We propose a unified adaptive inference framework that combines Token Reduction (TR) and Early Exiting (EE) through dataset-specific profiling. Our approach quantifies spatial redundancy via Jensen-Shannon Divergence (JSD) and prediction confidence at intermediate layers to train a lightweight predictor that dynamically selects inference strategies at test time. Across five medical datasets, including a real-world cataract dataset (INSIGHT), our framework achieves 71.4% average floating-point operations per second (FLOPs) reduction with only 0.1pp accuracy loss, substantially outperforming individual strategies (EE-only: 55.9%, TR-only: 57.7%). On PathMNIST, our adaptive inference framework simultaneously improves accuracy by 1.3pp while reducing computation by 77.2%. On INSIGHT, we maintain baseline accuracy with 69.8% FLOPs reduction, demonstrating robust real-world clinical applicability.

Keywords: Vision Transformers, Efficient Inference, Token Reduction, Early Exiting.

* Contributed equally
† Corresponding authors

1. Introduction

Vision Transformers (ViTs) have achieved state-of-the-art performance across medical imaging tasks, including dermatological lesion classification (Himel et al., 2024; Al-Waisy et al., 2025), chest X-ray diagnosis (Singh et al., 2024), histopathological tissue analysis (Xu et al., 2023b), and ophthalmic image analysis (Wu et al., 2023), leveraging self-attention to capture long-range dependencies crucial for complex diagnostic tasks (Dosovitskiy et al., 2021). However, clinical deployment faces critical computational barriers in time-sensitive and resource-constrained settings. High-volume screening programs for diabetic retinopathy or cataract screening must process thousands of images daily (Ruamviboonsuk et al., 2022; Tham et al., 2022), where even modest per-image latency accumulates into substantial computational burden. Point-of-care imaging on mobile devices (Xu et al., 2025) operates under severe hardware constraints, making standard ViT models impractical for these scenarios.

Approaches to address computational inefficiency in deep learning divide into model-centric optimizations (quantization (Li and Gu, 2023; Du et al., 2024), compressin (Wang et al., 2022b; Zhang et al., 2022), efficient attention (Han et al., 2023)) that apply static architectural changes, and data-centric methods that dynamically adapt to input characteristics. Two prominent data-centric strategies are TR (Rao et al., 2021; Liang et al., 2022; Bolya et al., 2022), which eliminates uninformative tokens, and EE (Bakhtiarnia et al., 2022; Xu et al., 2023a), which terminates inference when samples achieve sufficient confidence. These data-centric approaches are particularly well-suited for medical imaging, where substantial variability exists both across and within modalities (Kline et al., 2022). In safety-critical medical applications, matching the appropriate strategy to dataset-specific characteristics is crucial, as suboptimal choices risk compromised diagnostic accuracy.

To address this gap, we introduce a unified framework for adaptive strategy selection across diverse medical imaging datasets (ISIC2019, PathMNIST, PneumoniaMNIST, RetinaMNIST, INSIGHT). We calibrate dataset-specific thresholds: TR threshold via Jensen-Shannon Divergence between attention distributions, and EE confidence thresholds at checkpoints. At inference, a lightweight CNN predictor estimates redundancy from input images to activate TR while intermediate heads enable early termination based on confidence. This ensures redundant samples undergo TR, high-confidence cases EE, and complex cases receive full processing, maximizing efficiency without compromising diagnostic accuracy.

Our main contributions are:

- **Unified Adaptive Framework:** We propose a unified framework that integrates TR and EE for ViTs, utilizing a lightweight predictor to adaptively activate TR based on input-specific spatial redundancy while leveraging confidence-based EE at intermediate layers, enabling instance-level optimization of both spatial and temporal redundancy.
- **Dataset-Specific Profiling Methodology:** We conduct comprehensive profiling analysis to characterize redundancy-complexity profiles through token-level similarity and sample-wise confidence distribution, revealing that optimal strategies vary across datasets.

- **Superior Efficiency-Accuracy Trade-offs:** Through comprehensive evaluation across five medical imaging datasets, our unified framework achieves 71.4% average FLOPs reduction with 0.1pp average accuracy degradation, outperforming individual strategies (EE-only: 55.9%, TR-only: 57.7%).

2. Methodology

This study proposes an integrated framework combining TR and EE to reduce computational cost in ViTs by adapting TR globally based on input image characteristics while using local prediction confidence for EE decisions. Our approach consists of three stages: (1) fine-tuning Data-efficient Image Transformer-Small (DeiT-S) (Touvron et al., 2021) with EE heads on the training set, (2) profiling the validation set to calibrate TR and EE thresholds, and (3) deploying the unified strategy at test-time inference. Detailed dataset statistics are provided in Table A1.

2.1. Stage 1: Model Training with Early Exit Heads

We fine-tune DeiT-S with 12 transformer blocks, attaching lightweight classifier heads at layers 4, 7, and 10. Each EE head operates on the CLS token using a two-layer MLP with layer normalization. We train all heads simultaneously using a weighted multi-exit loss:

$$L_{\text{total}} = w_{\text{final}} \cdot L_{\text{final}} + \sum_{k \in \{4,7,10\}} w_k \cdot L_k \tag{1}$$

where L_k is the cross-entropy loss at layer k. We set blue$w_4 = w_7 = w_{10} = 0.3$, and $w_{\text{final}} = 1.0$.

2.2. Stage 2: Dataset-Specific Profiling for Strategy Selection

We use the validation set to profile dataset-specific characteristics and calibrate thresholds for adaptive inference: (1) EE confidence thresholds θ_{EE} for each checkpoint, (2) ground truth redundancy scores for training the lightweight predictor, and (3) the redundancy threshold θ_{R} for TR activation.

2.2.1. Early Exit Threshold Calibration

We calibrate dataset-specific thresholds θ_{EE} by sweeping confidence values on the validation set, selecting thresholds that maximizes FLOPs reduction while maintaining accuracy degradation <1%. At inference, when confidence $c_k = \max(\text{Softmax}(\mathbf{z}_k))$ at layer k exceeds θ_{EE}, the sample is classified and inference terminates. Otherwise, computation continues to the next block.

2.2.2. Spatial Redundancy Profiling

To determine which samples benefit from TR, we quantify spatial redundancy using attention similarity. For each validation sample, we compute the ground truth redundancy score y_{red} based on JSD between attention distributions:

$$y_{\text{red}} = 1 - \frac{1}{3}\left(\text{JSD}(\mathbf{a}_1, \mathbf{a}_4) + \text{JSD}(\mathbf{a}_4, \mathbf{a}_7) + \text{JSD}(\mathbf{a}_7, \mathbf{a}_10)\right) \tag{2}$$

where $\mathbf{a}_i$ represents the attention distribution at layer i. We select layers to balance coverage and efficiency: early layers alone miss semantic patterns, late layers overlook initial redundancy, while sparser/denser sampling either misses dynamics or adds unnecessary overhead. High y_{red} values (close to 1) indicate low divergence between attention patterns across layers, suggesting high spatial redundancy where tokens can be safely reduced.

2.2.3. Lightweight Redundancy Predictor Training

To avoid the computational overhead of full forward passes at test time, we train a lightweight custom CNN predictor to estimate redundancy from input images. The `Score_Predictor` consists of a feature extractor with three convolutional blocks. The blocks progressively increase channel depth ($32 \to 64 \to 128$) using 3×3 convolutions (stride 2), followed by Batch Normalization and ReLU activation. It predicts $\hat{y}_{\text{red}} \in [0, 1]$ from input image $\mathbf{x} \in \mathbb{R}^{H \times W \times C}$ using mean squared error loss:

$$\mathcal{L}_{\text{pred}} = \frac{1}{N} \sum_{i=1}^{N} (\hat{y}_{\text{red}}^{(i)} - y_{\text{red}}^{(i)})^2 \tag{3}$$

2.2.4. Token Reduction Threshold Calibration

We establish the TR activation threshold θ_{R} by analyzing the distribution of predicted redundancy scores on the separate validation set. For each validation sample, we first calculate accuracy–FLOPs pairs under identical EE settings for two scenarios: one without TR and one with TR applied at every checkpoint. We then sweep candidate thresholds, using the predicted redundancy score $\hat{y}_{\text{red}}$ to assign each sample to one of the two paths, and sum the previously calculated values to estimate overall accuracy and cost. We select the smallest threshold that maintains accuracy within 1% of the all-token baseline while maximizing FLOPs reduction. At test time, if $\hat{y}_{\text{red}} > \theta_{\text{R}}$, TR is activated after each EE checkpoint (layers 4, 7, and 10).

2.3. Stage 3: Unified Inference at Test Time

The complete inference pipeline for each test sample:

1. The `Score_Predictor` estimates $\hat{y}_{\text{red}}$ and sets TR activation flag: `use_tr` $= \hat{y}_{\text{red}} > \theta_{\text{R}}$
2. The ViT processes the image layer-by-layer through the 12 transformer blocks
3. At each checkpoint (layers 4, 7, 10):
   - The EE head computes confidence $c_k = \max(\text{Softmax}(\mathbf{z}_k))$
   - If $c_k > \theta_{\text{EE},k}$, inference terminates and returns $\arg\max(\mathbf{z}_k)$
   - If TR is activated (`use_tr = True`) and $c_k \leq \theta_{\text{EE},k}$, apply TR before continuing
4. If no EE occurs, the final head at layer 12 produces the classification

This unified pipeline enables dataset-adaptive optimization: spatially redundant samples undergo TR,high-confidence samples EE, and challenging samples process through all layers. Algorithm 1 formalizes this complete inference procedure.

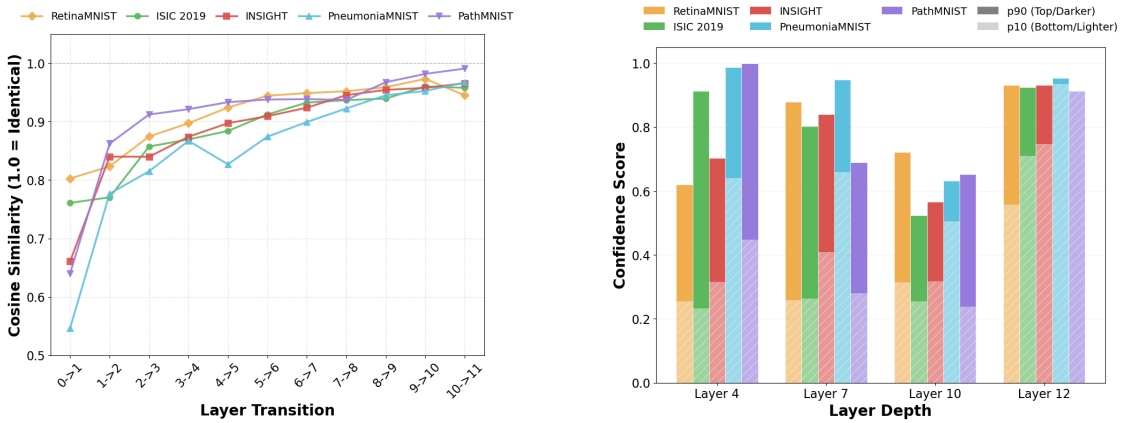
3. Results and Discussion

3.1. Experimental Setup

We evaluate our framework across five public datasets: ISIC2019 (Tschandl et al., 2018; Codella et al., 2018; Hernández-Pérez et al., 2024) (7-class skin lesion classification), PathMNIST (Yang et al., 2023) (9-class colon tissue classification), PneumoniaMNIST (2-class pneumonia detection), RetinaMNIST (5-class diabetic retinopathy grading), and INSIGHT, a private dataset of anterior segment eye images (4-class cataract classification). All images were resized to a 224×224 pixels using bicubic interpolation. For training, we applied data augmentation: random affine transformations (rotation range $\pm 10°$, translation up to 10%), autocontrast ($p = 0.5$), and horizontal flipping. For testing and validation, images were resized to 224×224. DeiT-S (Touvron et al., 2021) trained with AdamW optimizer (Loshchilov and Hutter, 2017) (learning rate: 5×10^{-4}, batch size: 64, epochs: 50).

3.2. Dataset Redundancy and Complexity Analysis

We profile dataset-specific redundancy using token-level cosine similarity across DeiT-S layers (Figure 1(a)). RetinaMNIST and ISIC2019 exhibit high initial similarity (∼0.8), while INSIGHT, PathMNIST, and PneumoniaMNIST start at moderate similarity (∼0.6). All datasets show monotonically increasing similarity with depth, confirming progressive feature homogenization (Zhou et al., 2021; Wang et al., 2022a).



(a) (b)

Figure 1: **Dataset redundancy and complexity analysis across DeiT-S layers.** (a) Token-level redundancy across layer transitions. RetinaMNIST and ISIC2019 show high initial similarity (∼0.8); others start lower (∼0.6). All exhibit monotonically increasing similarity. (b) Sample-wise complexity at decision layers (4, 7, 10, 12) showing 90th percentile (easy, lighter) and 10th percentile (hard, darker) confidence. PathMNIST and PneumoniaMNIST achieve high early confidence with minimal easy-hard gaps. RetinaMNIST and INSIGHT show persistent gaps.

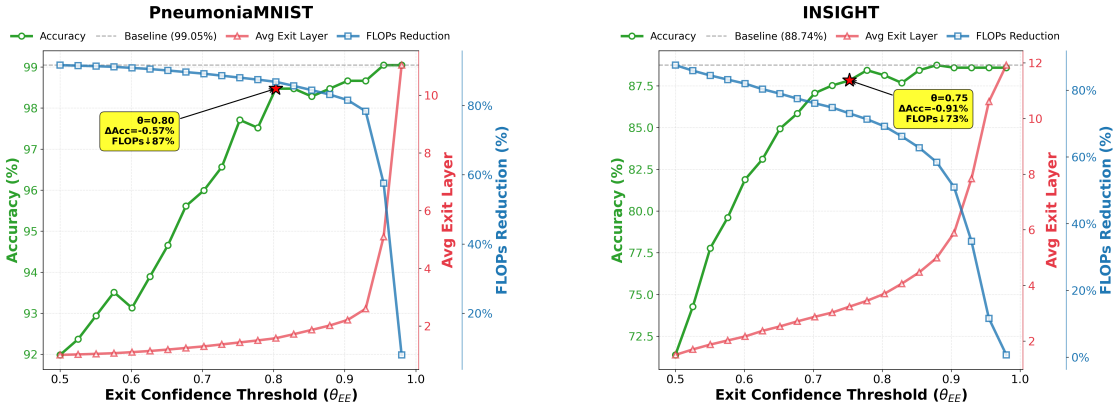
Figure 1(b) shows layer-wise confidence evolution at decision layers (4, 7, 10, 12). PathMNIST and PneumoniaMNIST achieve high early confidence (>0.8 by layer 4) with minimal

easy-hard gaps at layer 12, indicating uniform sample complexity. Conversely, RetinaMNIST and INSIGHT start with low confidence and maintain substantial easy-hard gaps through layer 12, reflecting diverse sample complexity. These profiles necessitate adaptive strategy selection: datasets with high early confidence and small gaps (PathMNIST, PneumoniaMNIST) suit aggressive EE, while those with persistent gaps (RetinaMNIST, INSIGHT) benefit more from TR or conservative thresholds.

3.3. Dataset-Specific Profiling for Strategy Selection

3.3.1. Early Exit Threshold Calibration

To determine dataset-specific optimal EE thresholds for subsequent experiments, we perform validation set profiling by sweeping $\theta_{\text{EE}} \in [0.5, 0.95]$ to maximize FLOPs reduction while constraining accuracy loss to $< 1\%$. Figure 2 illustrates the performance-efficiency trade-offs for PneumoniaMNIST and INSIGHT (remaining datasets in Figure A1).



(a) (b)

Figure 2: **Performance-efficiency trade-offs of early exiting.** Accuracy (green), FLOPs reduction (blue), and exit layer (pink) vs. confidence threshold. (a) PneumoniaMNIST achieves 87% FLOPs reduction at $\theta_{\text{EE}} = 0.80$ (average exit layer 1.58, -0.57pp from baseline). (b) INSIGHT requires $\theta_{\text{EE}} = 0.75$ (average exit layer 3.24, -0.61pp from baseline) for 73% FLOPs reduction.

PneumoniaMNIST (Figure 2(*a*)) achieves 87% FLOPs reduction at $\theta_{\text{EE}} = 0.80$ (average exit layer 5.8, -0.57pp from baseline). INSIGHT (Figure 2(*b*)) requires $\theta_{\text{EE}} = 0.75$, exiting at layer 10.6 for 73% reduction (-0.61pp from baseline). This disparity confirms that datasets with high early-layer similarity enable aggressive EE, while those with diverse initial representations require deeper inference. Based on this profiling, we select: $\theta_{\text{EE}}^{\text{Retina}} = 0.78$, $\theta_{\text{EE}}^{\text{Pneumonia}} = 0.80$, $\theta_{\text{EE}}^{\text{INSIGHT}} = 0.75$, $\theta_{\text{EE}}^{\text{ISIC}} = 0.65$, $\theta_{\text{EE}}^{\text{Pathology}} = 0.60$.

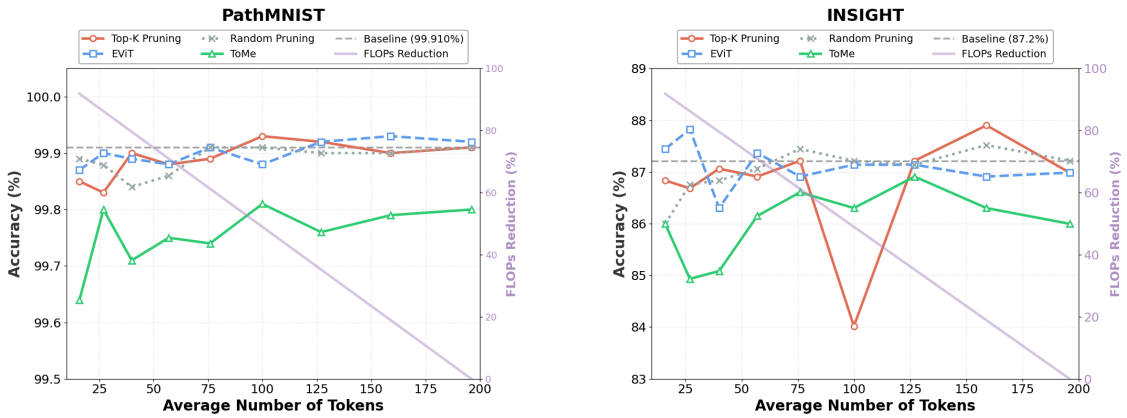
3.3.2. Token Reduction Keep Rate Selection

Figure 3 presents the accuracy-efficiency trade-offs of various TR strategies—random pruning, Top-K pruning, EViT (Liang et al., 2022), and Token Merging (ToMe) (Bolya et al.,

2022)—as a function of average token count across all DeiT-S layers, revealing substantial differences in robustness across datasets.

PathMNIST demonstrates resilience (Figure 3(*a*)): all strategies maintain performance above 99.6%, even with aggressive reduction to 16 tokens (ToMe) versus the 99.91% baseline. Random, Top-K, and EViT maintain ∼99.9% accuracy across all token budgets. INSIGHT shows distinct sensitivity (Figure 3(*b*)): while Random pruning and EViT preserve ∼87% accuracy down to 76 tokens, Top-K drops sharply to 84.0% at 100 tokens (3.2% loss from the 87.2% baseline). This validates Figure 1(*a*): PathMNIST's higher token similarity indicates more redundant spatial information, which is precisely what TR exploits: when tokens are similar, fewer are needed to preserve discriminative features.

Validation set profiling identifies EViT as the most stable strategy with minimal sensitivity to token budgets (Figure A2). The optimal keep rates (the proportion of tokens preserved at each layer) vary: ISIC2019/RetinaMNIST: 0.3, PneumoniaMNIST: 0.4, INSIGHT: 0.5, PathMNIST: 0.7. For fair comparison across TR-only and TR+EE configurations, we standardize at keep rate 0.4 for all experiments. At 40 retained tokens, EViT delivers a consistent ∼56% FLOPs reduction (2.027 vs. 4.608 GFLOPs baseline) with competitive accuracy: PathMNIST 99.89% (−1.2pp from baseline), PneumoniaMNIST 97.9% (+2.0pp), INSIGHT 86.3% (−0.9pp), RetinaMNIST 60.0% (+1.0pp), ISIC2019 56.78% (+2.6pp).



(*a*) (*b*)

Figure 3: **Token reduction strategy comparison across medical imaging datasets.** X-axis: average token count; y-axis: accuracy (%, left) and FLOPs reduction (%, right, purple). Methods: Top-K (red circles), EViT (blue squares), Random (gray crosses), ToMe (green triangles). Gray dashed line: baseline accuracy. (a) PathMNIST: All strategies maintain > 99.7% accuracy at 40 tokens; EViT achieves 99.89% (−1.2pp). (b) INSIGHT: EViT shows superior stability, maintaining 86.3% at 40 tokens (−0.9pp), while Top-K and Random exhibit higher variance. ToMe collapses below 76 tokens.

3.4. Lightweight Redundancy Predictor Performance

Table 1 demonstrates that the lightweight `Score_Predictor` effectively approximates spatial redundancy. RetinaMNIST, PathMNIST, and INSIGHT achieve strong performance (MAE ≤ 0.082, Pearson $R \geq 0.70$), while ISIC2019 and PneumoniaMNIST show weaker correlations (Pearson $R = 0.48$ and 0.31 due to narrower dynamic ranges in their redundancy distributions. Nevertheless, the predictor remains sufficiently accurate for coarse separation between low- and high-redundancy samples to trigger TR decisions. The calibrated thresholds reveal dataset-specific regimes: RetinaMNIST exhibits the highest threshold ($\theta_{\mathrm{R}} = 0.9431$) to preserve subtle vascular patterns, while PathMNIST and INSIGHT adopt permissive thresholds ($\theta_{\mathrm{R}} = 0.0870$ and 0.1883) consistent with higher baseline redundancy, enabling dataset-aware efficiency without expensive redundancy estimation at test time.

Table 1: **Lightweight redundancy predictor performance on validation data.** The `Score_Predictor` estimates spatial redundancy scores to determine TR activation. Mean absolute error (MAE), Pearson R, and R^2 evaluate prediction performance. Optimal thresholds θ_{R} are calibrated per dataset to balance computational savings and accuracy.

Dataset	**MAE (↓)**	**Pearson R (↑)**	**$\mathbf{R^2}$ (↑)**	**Optimal Threshold ($\theta_{\mathbf{R}}$)**
ISIC 2019	0.1352	0.4756	0.2124	0.1952
PneumoniaMNIST	0.1631	0.3055	0.0730	0.2988
RetinaMNIST	0.0804	0.7917	0.3567	0.9431
PathMNIST	0.0635	0.7039	0.4701	0.0870
INSIGHT	0.0818	0.7408	0.4932	0.1883

3.5. Unified Framework Performance

Table 2 compares our unified TR+EE framework against baseline and individual strategies across five datasets. Our approach achieves 71.4% average FLOPs reduction while maintaining accuracy, substantially outperforming individual strategies (EE-only: 55.9%, TR-only: 55.7%).

The results highlight distinct behaviors across datasets, validating the need for dataset-specific profiling. PathMNIST achieves 96.0% accuracy (+1.3pp) and 77.2% FLOPs reduction, as TR removes redundant background tokens enabling EE to focus on salient tissue structures and exit early (avg. layer 3.0). Similarly, on the real-world INSIGHT dataset maintains baseline accuracy (86.2%) with 69.8% FLOPs reduction, exploiting the high spatial redundancy typical of anterior segment imaging (avg. 29.2 tokens).

However, ISIC2019 and PneumoniaMNIST show minor degradation. ISIC2019's 2.5pp loss stems from aggressive TR (20.8 tokens, 10% of original) at layer 10.3, eliminating subtle diagnostic features. PneumoniaMNIST's 1.3pp loss indicates premature termination on subtle cases. While TR-only improved accuracy (92.1%), the unified framework's aggressive EE prevents token-reduced representations from reaching deeper layers where they could recover performance. Both cases demonstrate the inherent trade-off: substantial computational savings (66.0% and 73.9% reduction) require accepting minor accuracy loss.

Table 2: **Unified Framework Performance Across Datasets.** All methods use DeiT-S backbone. Baseline: 196 tokens, 12 layers. EE-only: 196 tokens with dynamic EE (dataset-specific θ_{EE}). TR-only: EViT with 40 tokens across all layers. TR+EE (Ours, shaded): combines TR and EE. Best results in **bold**.

Dataset	Strategy	Accuracy (%)	Avg Tokens	Avg Exit Layer	FLOPs (G)
ISIC2019	Baseline	54.2	196	12.0	4.61
	EE-only	**56.8** (↑2.6pp)	196	**9.26**	2.616 (↓43.3%)
	TR-only	**56.8** (↑2.6pp)	40	12.0	2.027 (↓56.0%)
	TR+EE	51.7 (↓2.5pp)	**20.8**	10.3	**1.569** (↓**66.0%**)
PneumoniaMNIST	Baseline	90.1	196	12.0	4.61
	EE-only	87.8 (↓2.3pp)	196	**1.86**	**0.776** (↓**83.2%**)
	TR-only	**92.1** (↑2.0pp)	**40**	12.0	2.027 (↓56.0%)
	TR+EE	88.8 (↓1.3pp)	56.3	4.65	1.204 (↓73.9%)
RetinaMNIST	Baseline	59.0	196	12.0	4.61
	EE-only	**61.0** (↑2.0pp)	196	**5.45**	1.662 (↓63.9%)
	TR-only	54.5 (↓4.5pp)	**40**	12.0	2.027 (↓56.0%)
	TR+EE	60.8 (↑1.8pp)	44.8	7.7	**1.398** (↓**70.0%**)
PathMNIST	Baseline	94.7	196	12.0	4.61
	EE-only	94.4 (↓0.3pp)	196	11	3.049 (↓33.9%)
	TR-only	93.5 (↓1.2pp)	**40**	12.0	2.027 (↓56.0%)
	TR+EE	**96.0** (↑**1.3pp**)	79	**3.0**	**1.05** (↓**77.2%**)
INSIGHT	Baseline	86.1	196	12.0	4.61
	EE-only	87.4 (↑1.3pp)	196	7.04	2.061 (↓55.3%)
	TR-only	85.5 (↓0.6pp)	40	12.0	1.633 (↓64.6%)
	TR+EE	**86.2** (↑**0.1pp**)	**29.2**	**6.9**	**1.394** (↓**69.8%**)
Average Performance Across All Datasets					
	EE-only	−0.3pp	196	7.0	2.033 (↓55.9%)
	TR-only	−0.4pp	**40**	12.0	1.948 (↓57.7%)
	TR+EE	**-0.1pp**	46.0	**6.5**	**1.323** (↓**71.4%**)

Per-class analysis (Table A2) reveals that our framework can improve sensitivity beyond baseline for diagnostically challenging classes, such as PathMNIST's mucus (98.16%) and cancer-associated stroma (72.92%), demonstrating that adaptive inference enhances performance on complex tissue types.

We conducted an ablation study on backbone architecture by evaluating ViT-S (Table A3). TR+EE achieves 54.1% to 76.4% FLOPs reduction across five datasets with accuracy changes ranging from -4.5pp to +0.2pp relative to baseline. These results suggest that architectural characteristics influence the efficiency-accuracy trade-off, informing model selection for clinical deployment with varying computational and accuracy requirements.

3.6. Visualization

Figure 4 visualizes the adaptive framework behavior on representative samples from INSIGHT and PathMNIST datasets. Figure 4(*a*) demonstrates sequential TR across all checkpoints with 40% retention rate (keep rate = 0.4), where the model processes all layers for

prediction while successfully preserving informative regions around the pupil even at the final checkpoint. Figure 4(*b*) shows a case where TR activates at the first checkpoint followed by EE, as the clear evidence of mature cataract at the pupil enables confident prediction without deeper processing. Similarly, Figure 4(*c*) illustrates a PathMNIST sample where TR activation at the first checkpoint is followed by EE, demonstrating the framework's ability to adaptively combine both efficiency strategies based on sample characteristics.

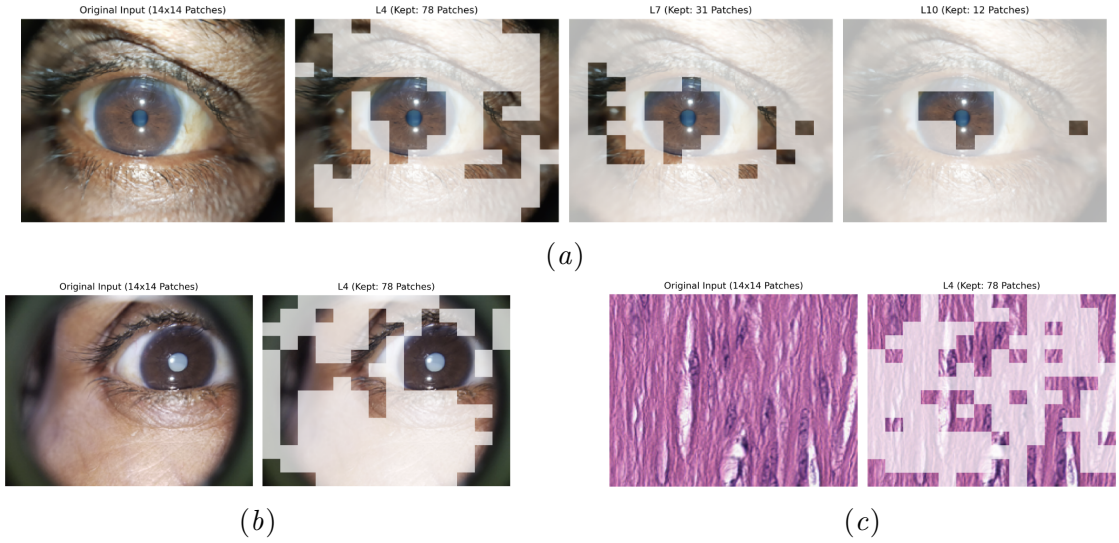


Figure 4: **Visualization of adaptive framework on medical imaging samples.** (a) TR-only with sequential token reduction maintaining diagnostic regions across all layers. (b), (c) Combined TR+EE examples from INSIGHT and PathMNIST.

4. Conclusion

This work introduces a unified framework that integrates TR and EE for ViT inference in medical imaging. We calibrate dataset-specific thresholds: prediction confidence for EE (θ_{EE}) and spatial redundancy via JSD for TR (θ_{R}). At test time, a lightweight CNN predictor estimates sample-level redundancy to activate TR, while intermediate classifier heads enable EE based on confidence. Across five diverse datasets, our framework achieves 71.4% average FLOPs reduction while maintaining diagnostic accuracy within 0.1pp of baseline, substantially outperforming individual strategies (EE-only: 55.9%; TR-only: 57.7%).

Clinical Impact: Medical AI deployment demands both accuracy and efficiency, particularly in resource-constrained and time-sensitive settings. Our framework achieves substantial efficiency gains without compromising performance on both public benchmarks and real-world clinical data with inherent quality variability. This approach can enable broader access to diagnostic AI where hardware resources are scarce but patient need is greatest.

Limitations: Our framework requires dataset-specific profiling to calibrate thresholds, but this one-time overhead maximizes test-time efficiency without recurring costs. Beyond theoretical FLOPs reductions, validation on edge devices is necessary to assess actual latency improvements.

Acknowledgments

We acknowledge support from the National Eye Institute (National Eye Institute (P30EY001765, R21EY034343), VentureWell Propel Award, Microsoft Acceleration Award, Stephen F Raab and Mariellen Brickley-Raab Rising Professorship in Ophthalmology, Johns Hopkins University, and the National Academy of Medicine. In addition, funds to support this AITC study were provided by the Johns Hopkins University AITC under award number P30AG073104.

References

Alaa S Al-Waisy, Shumoos Al-Fahdawi, Mohammed I Khalaf, Mazin Abed Mohammed, Bourair Al-Attar, and Mohammed Nasser Al-Andoli. A deep learning framework for automated early diagnosis and classification of skin cancer lesions in dermoscopy images. *Scientific Reports*, 15(1):31234, 2025.

Arian Bakhtiarnia, Qi Zhang, and Alexandros Iosifidis. Single-layer vision transformers for more accurate early exits with less overhead. *Neural Networks*, 153:461–473, 2022.

Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your vit but faster. *arXiv preprint arXiv:2210.09461*, 2022.

Noel CF Codella, David Gutman, M Emre Celebi, Brian Helba, Michael A Marchetti, Stephen W Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, et al. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*, pages 168–172. IEEE, 2018.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. URL https://arxiv.org/abs/2010.11929.

Dayou Du, Gu Gong, and Xiaowen Chu. Model quantization and hardware acceleration for vision transformers: A comprehensive survey. *arXiv preprint arXiv:2405.00314*, 2024.

Dongchen Han, Xuran Pan, Yizeng Han, Shiji Song, and Gao Huang. Flatten transformer: Vision transformer using focused linear attention. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5961–5971, 2023.

Carlos Hernández-Pérez, Marc Combalia, Sebastian Podlipnik, Noel CF Codella, Veronica Rotemberg, Allan C Halpern, Ofer Reiter, Cristina Carrera, Alicia Barreiro, Brian Helba, et al. Bcn20000: Dermoscopic lesions in the wild. *Scientific data*, 11(1):641, 2024.

Galib Muhammad Shahriar Himel, Md Masudul Islam, Kh Abdullah Al-Aff, Shams Ibne Karim, and Md Kabir Uddin Sikder. Skin cancer segmentation and classification using vision transformer for automatic analysis in dermatoscopy-based noninvasive digital system. *International Journal of Biomedical Imaging*, 2024(1):3022192, 2024.

Adrienne Kline, Hanyin Wang, Yikuan Li, Saya Dennis, Meghan Hutch, Zhenxing Xu, Fei Wang, Feixiong Cheng, and Yuan Luo. Multimodal machine learning in precision health: A scoping review. *NPJ digital medicine*, 5(1):171, 2022.

Zhikai Li and Qingyi Gu. I-vit: Integer-only quantization for efficient vision transformer inference. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17065–17075, 2023.

Youwei Liang, Chongjian Ge, Zhan Tong, Yibing Song, Jue Wang, and Pengtao Xie. Not all patches are what you need: Expediting vision transformers via token reorganizations. *arXiv preprint arXiv:2202.07800*, 2022.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Yongming Rao, Wenliang Zhao, Benlin Liu, Jiwen Lu, Jie Zhou, and Cho-Jui Hsieh. Dynamicvit: Efficient vision transformers with dynamic token sparsification. *Advances in neural information processing systems*, 34:13937–13949, 2021.

Paisan Ruamviboonsuk, Richa Tiwari, Rory Sayres, Variya Nganthavee, Kornwipa Hemarat, Apinpat Kongprayoon, Rajiv Raman, Brian Levinstein, Yun Liu, Mike Schaekermann, et al. Real-time diabetic retinopathy screening by deep learning in a multisite national screening programme: a prospective interventional cohort study. *The Lancet Digital Health*, 4(4):e235–e244, 2022.

Sukhendra Singh, Manoj Kumar, Abhay Kumar, Birendra Kumar Verma, Kumar Abhishek, and Shitharth Selvarajan. Efficient pneumonia detection using vision transformers on chest x-rays. *Scientific reports*, 14(1):2487, 2024.

Yih-Chung Tham, Jocelyn Hui Lin Goh, Ayesha Anees, Xiaofeng Lei, Tyler Hyungtaek Rim, Miao-Li Chee, Ya Xing Wang, Jost B Jonas, Sahil Thakur, Zhen Ling Teo, et al. Detecting visually significant cataract using retinal photograph-based deep learning. *Nature aging*, 2(3):264–271, 2022.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers distillation through attention, 2021. URL https://arxiv.org/abs/2012.12877.

Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, 5(1):1–9, 2018.

Peihao Wang, Wenqing Zheng, Tianlong Chen, and Zhangyang Wang. Anti-oversmoothing in deep vision transformers via the fourier domain analysis: From theory to practice. *arXiv preprint arXiv:2203.05962*, 2022a.

Zhenyu Wang, Hao Luo, Pichao Wang, Feng Ding, Fan Wang, and Hao Li. Vtc-lfc: Vision transformer compression with low-frequency components. *Advances in Neural Information Processing Systems*, 35:13974–13988, 2022b.

Jo-Hsuan Wu, Neslihan D Koseoglu, Craig Jones, and TY Alvin Liu. Vision transformers: The next frontier for deep learning-based ophthalmic image analysis. *Saudi Journal of Ophthalmology*, 37(3):173–178, 2023.

Guanyu Xu, Jiawei Hao, Li Shen, Han Hu, Yong Luo, Hui Lin, and Jialie Shen. Lgvit: Dynamic early exiting for accelerating vision transformer. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 9103–9114, 2023a.

Hongming Xu, Qi Xu, Fengyu Cong, Jeonghyun Kang, Chu Han, Zaiyi Liu, Anant Madabhushi, and Cheng Lu. Vision transformers for computational histopathology. *IEEE Reviews in Biomedical Engineering*, 17:63–79, 2023b.

Yiwen Xu, Tariq M Khan, Yang Song, and Erik Meijering. Edge deep learning in computer vision and medical diagnostics: a comprehensive survey. *Artificial Intelligence Review*, 58(3):93, 2025.

Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific Data*, 10(1):41, 2023.

Jinnian Zhang, Houwen Peng, Kan Wu, Mengchen Liu, Bin Xiao, Jianlong Fu, and Lu Yuan. Minivit: Compressing vision transformers with weight multiplexing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12145–12154, 2022.

Daquan Zhou, Bingyi Kang, Xiaojie Jin, Linjie Yang, Xiaochen Lian, Zihang Jiang, Qibin Hou, and Jiashi Feng. Deepvit: Towards deeper vision transformer, 2021. URL https://arxiv.org/abs/2103.11886.

Table A1: Dataset statistics and data split information. The validation set is split into two equal parts: 50% for dataset-specific profiling and lightweight predictor training, and 50% for token reduction threshold calibration.

Dataset	**Train**	**Validation**	**Test**
ISIC2019	1791	448	118
PneumoniaMNIST	4,708	524	624
RetinaMNIST	1,080	120	400
PathMNIST	89,996	10,004	7,180
INSIGHT	5256	657	657

INSIGHT Ethics Statement After obtaining informed consent, community health workers collect smartphone-based images of patients attending community eye screenings. Diagnosis labels for each image were obtained using clinical diagnoses made via pen light examination by ophthalmologists at the same screening. The study was approved by the Institutional Review Boards of Aravind Eye Hospital and the Johns Hopkins University School of Medicine.

Algorithm 1 Unified Adaptive Inference Pipeline (TR + EE)

Input: Input Image $\mathbf{x}$, Redundancy Threshold θ_{R}, EE Thresholds $\{\theta_{\text{EE}}\}$, Token Keep Rate r

Output: Prediction $\hat{y}$, Exit Layer k^*

```
// Stage 1: Adaptive Token Reduction (TR) Activation
```
$\hat{y}_{\text{red}} \leftarrow$ `Score_Predictor`$(\mathbf{x})$ `// Predict redundancy score`

`use_tr` $\leftarrow (\hat{y}_{\text{red}} > \theta_{\text{R}})$ `// Activate TR if redundant`

$\mathbf{Z} \leftarrow$ Initial Tokens with N_0 tokens

```
// Stage 2: Iterative Layer Processing with TR and EE
```
for $k \leftarrow 0$ **to** 11 **do**
  $\mathbf{Z} \leftarrow \text{Transformer\_Block}_k(\mathbf{Z})$
  if $k \in \{3, 6, 9\}$ **then**
    `// Early Exit Check`
    $\mathbf{z}_k \leftarrow \text{CLS\_Token\_Output}(\mathbf{Z})$ $c_k \leftarrow \max(\text{Softmax}(\mathbf{z}_k))$
    if $c_k > \theta_{EE}$ **then**
      return $\hat{y} \leftarrow \arg\max(\mathbf{z}_k)$, $k^* \leftarrow k$
    end
    `// Token Reduction (if active)`
    if *use_tr* $=$ *True* **then**
      $N_{\text{new}} \leftarrow N_{\text{current}} \cdot r$ $\mathbf{Z} \leftarrow \text{Reduce\_Tokens}(\mathbf{Z}, N_{\text{new}})$
    end
  end
end

```
// Stage 3: Default (Full Depth)
```
$\mathbf{z}_{12} \leftarrow \text{CLS\_Token\_Output}(\mathbf{Z})$ **return** $\hat{y} \leftarrow \arg\max(\mathbf{z}_{12})$, $k^* \leftarrow 12$

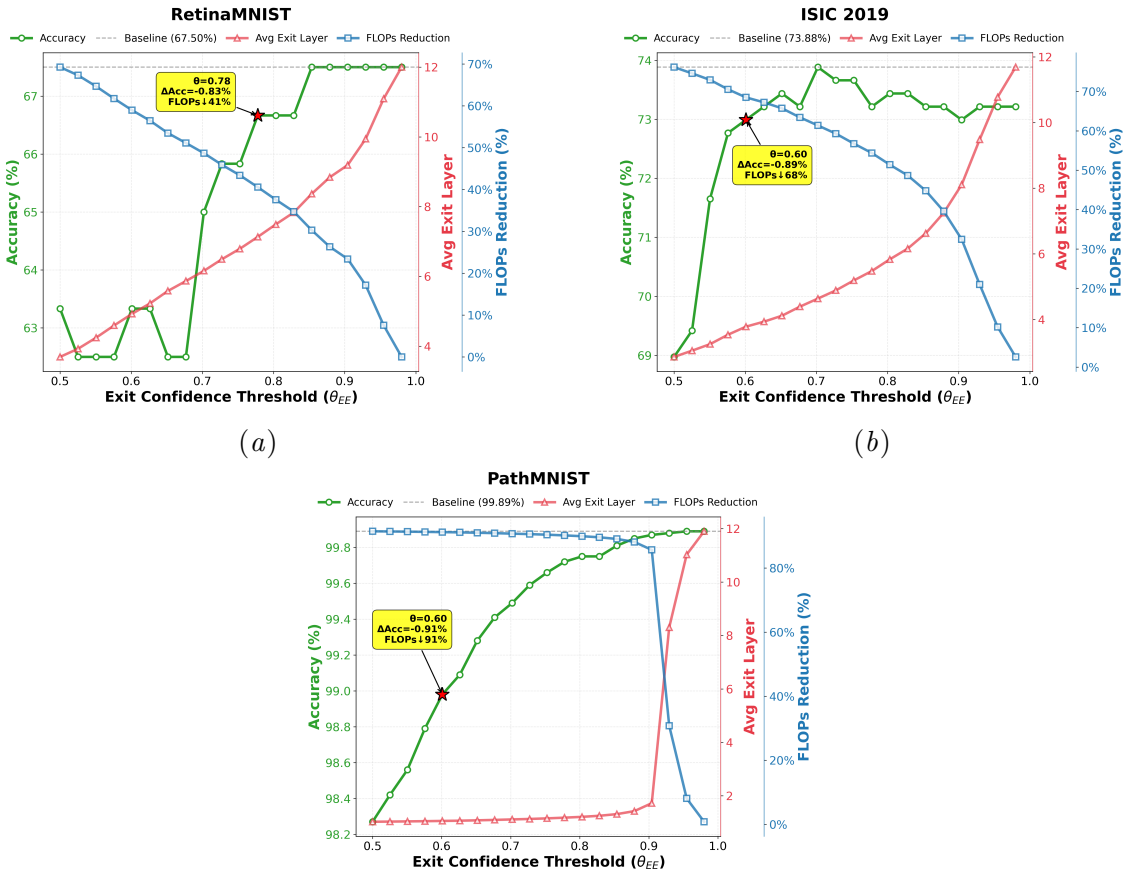


Figure A1: Performance-efficiency trade-offs of early exiting across confidence thresholds. Validation set profiling sweeps $\theta_{\text{EE}} \in [0.5, 0.95]$ to identify optimal thresholds maximizing FLOPs reduction while constraining accuracy loss to $< 1\%$. X-axis shows confidence threshold θ_{EE}; left y-axis shows accuracy (%, green circles), right y-axis shows FLOPs reduction (%, blue squares) and average exit layer (pink triangles). (a) RetinaMNIST: Achieves 41% FLOPs reduction at $\theta_{\text{EE}} = 0.78$ with 66.67% accuracy (-0.83pp from baseline). (b) ISIC2019: Achieves 65% FLOPs reduction at $\theta_{\text{EE}} = 0.65$ with 72.32% accuracy (-0.89pp from baseline). (c) PathMNIST: Achieves 91% FLOPs reduction at $\theta_{\text{EE}} = 0.60$ with 98.98% accuracy (-0.91pp from baseline)

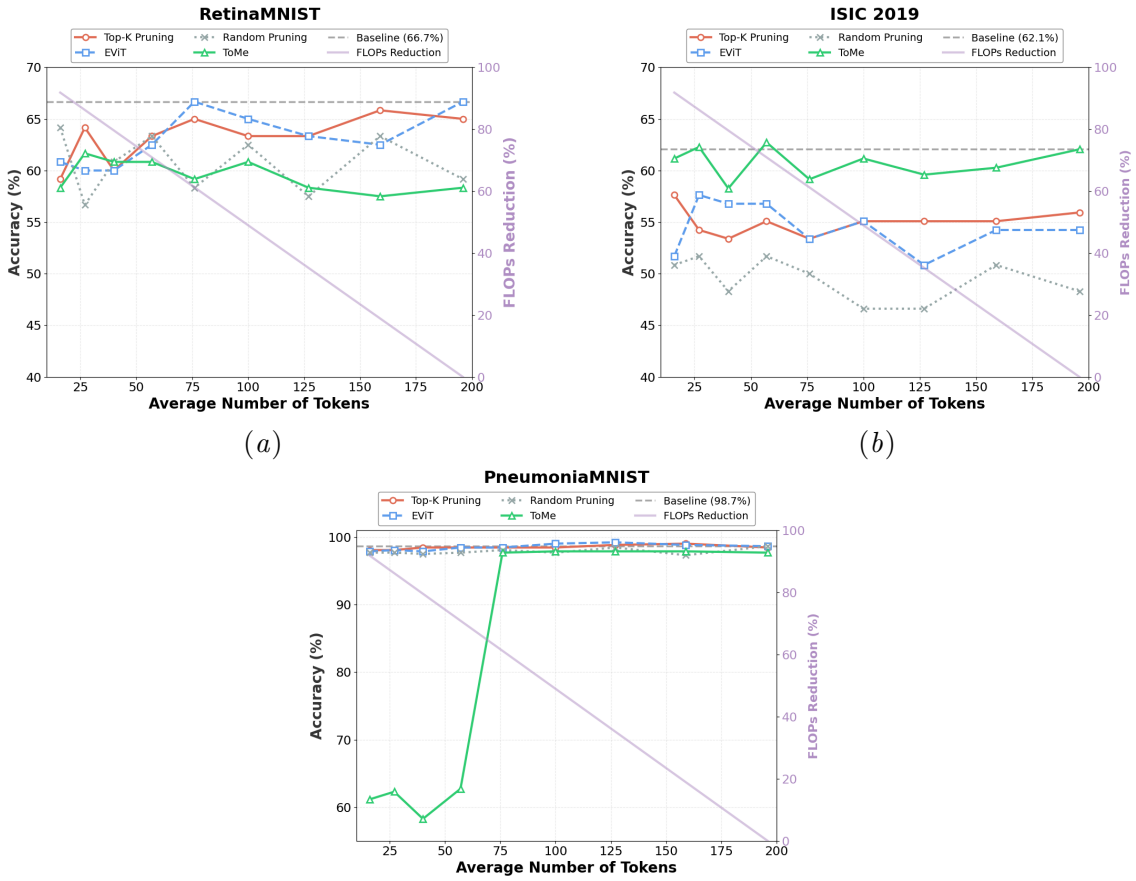


Figure A2: Token reduction strategy comparison across medical imaging datasets. X-axis shows average token count; y-axis shows accuracy (%, left) and FLOPs reduction (%, right, purple line). Methods: Top-K (red circles), EViT (blue squares), Random (gray crosses), ToMe (green triangles). Gray dashed line marks baseline accuracy. (a) RetinaMNIST: EViT maintains consistent performance (60-67% accuracy) across all token budgets. Top-K performs well at higher token counts (65.8% at 159 tokens), while ToMe shows degraded performance comparable to random pruning. (b) ISIC2019: High sensitivity to token reduction across all strategies. ToMe achieves best overall performance (peak 62.7% at 57 tokens). EViT excels at low token budgets while Top-K performs better at high token counts, showing complementary efficiency profiles. (c) PneumoniaMNIST: Highly robust to token reduction across all strategies, maintaining $> 97\%$ accuracy down to 16 tokens. EViT, Top-K, and Random show negligible degradation, while ToMe exhibits relative instability below 57 tokens despite maintaining strong absolute performance.

Table A2: **Diagnostic Performance and Latency by Strategy and Class.** Comparison of per-class Sensitivity (↑) and Specificity (↑) across five medical imaging datasets. The performance variations highlight dataset-specific biases. **Bolding** indicates the best metric across the four strategies for that specific class/metric. Latency is the achieved runtime (ms).

Dataset	Class Name	Sensitivity (%) (↑)				Specificity (%) (↑)			
		Baseline	EE-only	TR-only	TR+EE	Baseline	EE-only	TR-only	TR+EE
ISIC 2019	0: actinic keratosis	12.50	12.50	6.25	**31.25**	**100.00**	**100.00**	99.02	**100.00**
	1: basal cell carcinoma	87.50	87.50	87.50	**93.75**	94.12	95.14	**97.06**	87.25
	2: dermatofibroma	50.00	**56.25**	43.75	37.50	99.02	**100.00**	**100.00**	**100.00**
	3: melanoma	12.50	6.25	18.75	**18.75**	95.14	92.16	94.12	**96.08**
	4: nevus	87.50	**100.00**	**100.00**	87.50	72.55	73.53	71.57	**83.33**
	5: pigmented benign keratosis	81.25	**87.50**	**87.50**	81.25	89.22	**93.14**	91.18	82.35
	6: seborrheic keratosis	0.00	0.00	0.00	0.00	**100.00**	**100.00**	**100.00**	**100.00**
	7: squamous cell carcinoma	50.00	50.00	**56.25**	25.00	97.06	97.06	97.06	**98.04**
	8: vascular lesion	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	99.13	**100.00**	99.13
PneumoniaMNIST	0: normal	73.93	67.95	**80.77**	70.94	**99.74**	**99.74**	98.97	98.72
	1: pneumonia	**99.74**	**99.74**	98.97	98.72	73.93	67.95	**80.77**	70.94
RetinaMNIST	0: No DR	**89.66**	87.93	82.76	85.06	57.96	67.70	71.24	**73.45**
	1: Mild DR	0.00	17.39	**36.96**	0.00	98.59	96.61	85.88	**99.15**
	2: Moderate DR	41.38	40.22	31.52	**52.17**	86.04	**87.66**	88.64	80.84
	3: Severe DR	51.47	58.82	35.29	**70.59**	**95.48**	92.17	92.17	90.96
	4: Proliferative DR	**35.00**	30.00	20.00	15.00	98.42	98.16	98.42	**99.74**
PathMNIST	0: adipose	**99.33**	98.58	97.38	98.43	99.50	99.67	99.52	**99.91**
	1: background	**100.00**	**100.00**	**100.00**	**100.00**	99.42	99.07	99.27	**100.00**
	2: debris	95.87	97.94	99.12	**97.35**	**99.80**	99.56	**99.83**	99.66
	3: lymphocytes	**100.00**	**100.00**	**100.00**	98.90	99.88	99.48	98.75	**99.89**
	4: mucus	88.50	95.07	92.08	**98.16**	**99.72**	99.59	99.43	99.12
	5: smooth muscle	93.24	87.16	85.47	**92.57**	97.81	98.51	**98.66**	98.28
	6: normal colon mucosa	**97.98**	95.95	90.01	93.30	98.57	98.70	98.80	**99.72**
	7: cancer-associated stroma	65.80	66.51	67.70	**72.92**	**99.87**	99.73	99.39	**99.87**
	8: colorectal adenocarcinoma epithelium	96.76	94.16	95.94	**98.30**	99.48	**99.51**	99.08	99.08
INSIGHT	clear	88.22	90.45	82.64	**91.40**	87.17	86.88	**91.69**	83.67
	immature cataract	81.17	79.37	**87.22**	77.13	90.55	92.17	87.10	**93.09**
	mature cataract	80.77	**92.31**	82.69	83.08	99.68	99.52	**99.76**	99.68
	pciol	92.55	94.68	**97.34**	93.62	99.29	**99.82**	99.29	99.47

Table A3: **Unified Framework Performance Across Datasets Using ViT-S backbone.** Baseline: 196 tokens, 12 layers. EE-only: 196 tokens with dynamic EE (dataset-specific θ_{EE}). TR-only: EViT with 100 tokens across all layers. TR+EE (Ours, shaded): combines TR and EE. Best results in **bold**.

Dataset	Strategy	Accuracy (%)	Avg Tokens	Avg Exit Layer	FLOPs (G)
ISIC2019	Baseline	60.2	196	12.0	4.61
	EE-only	56.8 (↓3.4pp)	196	**3.75**	**1.236 (↓73.2%)**
	TR-only	**61.9** (↑1.7pp)	100	12.0	3.003 (↓34.8%)
	TR+EE	56.8 (↓3.4pp)	**68**	12.0	2.117 (↓**54.1%**)
PneumoniaMNIST	Baseline	94.2	196	12.0	4.61
	EE-only	93.9 (↓0.3pp)	196	**3.0**	**1.05 (↓77.2%)**
	TR-only	**95.0** (↑0.8pp)	**100**	12.0	3.003 (↓34.8%)
	TR+EE	90.4 (↓3.8pp)	110	5.53	1.429 (↓69.0%)
RetinaMNIST	Baseline	**66.8**	196	12.0	4.61
	EE-only	65.5 (↓1.3pp)	196	**6.77**	**1.992 (↓56.8%)**
	TR-only	63.0 (↓3.8pp)	100	12.0	3.003 (↓34.8%)
	TR+EE	62.3 (↓4.5pp)	**68**	12.0	2.117 (↓54.1%)
PathMNIST	Baseline	93.3	196	12.0	4.61
	EE-only	93.7 (↓0.5pp)	196	**3.13**	**1.082 (↓76.5%)**
	TR-only	**94.6** (↑0.4pp)	**100**	12.0	3.003 (↓34.8%)
	TR+EE	93.5 (↑0.2pp)	136	3.22	1.086 (↓76.4%)
INSIGHT	Baseline	86.9	196	12.0	4.61
	EE-only	**88.1** (↑1.2pp)	196	**3.38**	**1.145 (↓75.2%)**
	TR-only	87.7 (↑0.8pp)	100	12.0	3.003 (↓34.8%)
	TR+EE	86.0 (↓0.9pp)	**70**	9.46	1.979 (↓57.1%)