GI Endoscopy AI Diagnostic System

Machine Learning Training Report

Project:	Advanced Vision Transformer Ensemble for GI Endoscopy
Date:	November 11, 2025
Version:	1.0
Author:	AI Diagnosis Team

Executive Summary

This report documents the development and training of an advanced ensemble-based deep learning system for automated classification of gastrointestinal (GI) endoscopy images. The system employs state-of-the-art Vision Transformer (ViT) architectures, specifically DeiT3 and ViT Base, trained at 384×384 resolution with advanced data augmentation and optimization techniques.

Key Achievements

• Architecture: Ensemble of DeiT3 Small and ViT Base models

• Resolution: 384×384 pixels (high-resolution input)

• Advanced Techniques: MixUp augmentation, Focal Loss, Test-Time Augmentation (TTA), Label Smoothing

• Memory Optimization: Gradient accumulation, mixed precision training

• Performance: Optimized for medical image classification with class imbalance handling

1. Methodology

1.1 Dataset Structure

The training pipeline expects a directory structure with class-based folders. The system classifies 23 different GI conditions including Barrett's esophagus, esophagitis, polyps, ulcerative colitis, hemorrhoids, and anatomical landmarks.

1.2 Data Preprocessing

Images are resized to 384×384 pixels for high-resolution medical imaging. Normalization uses ImageNet statistics (mean: [0.485, 0.456, 0.406], std: [0.229, 0.224, 0.225]). All images are converted to RGB format and transformed to PyTorch tensors.

2. Model Architecture

2.1 Vision Transformer Architecture

Both models use the Vision Transformer architecture with 16×16 pixel patches, 384×384 input resolution, multi-head self-attention mechanism, and 23-class classification output.

3. Training Configuration

Parameter	Value	Rationale
Image Size	384×384	High resolution for medical detail
Batch Size	2	Memory constraints with 384px images
Effective Batch Size	16	Via gradient accumulation (8 steps)
Learning Rate	1e-5	Conservative learning rate for fine-tuning
Epochs	25	Sufficient for convergence
Optimizer	AdamW	Weight decay: 0.01
Weight Decay	0.01	Regularization

4. Advanced Training Techniques

MixUp Augmentation: 50% chance per batch, alpha=0.2, linear interpolation between images and labels

Gradient Accumulation: 8 accumulation steps, effective batch size of 16

Mixed Precision Training: FP16 for forward pass, FP32 for gradients, 2× memory reduction

Test-Time Augmentation: Original + horizontal flip + vertical flip, average of 3 predictions

Focal Loss: Alpha=1.0, Gamma=2.0, addresses class imbalance

Label Smoothing: Smoothing factor 0.1, prevents overconfidence

Cosine Warmup: 5 warmup epochs, then cosine annealing to 0

5. Data Augmentation

Training Augmentations:

• Resize: 384×384

• Horizontal Flip: p=0.5

• Vertical Flip: p=0.3

• Random Rotate 90°: p=0.5

• ShiftScaleRotate: ±10% shift/scale, ±15° rotation, p=0.5

• ColorJitter: ±20% brightness/contrast/saturation, ±10% hue, p=0.5

• Gaussian Noise: p=0.3

• CoarseDropout: Max 1 hole, 32×32 size, p=0.3

6. Code Architecture

6.1 Core Classes

FocalLoss: Custom loss function for class imbalance with alpha and gamma parameters

OptimizedMedicalDataset: PyTorch Dataset for medical images with automatic class discovery

AdvancedMemoryEfficientTrainer: Complete training pipeline with memory optimization

AdvancedMemoryEfficientEnsemble: Multi-model ensemble with weighted predictions

7. Computational Requirements

7.1 Hardware

Minimum: GPU with 8GB VRAM, 16GB RAM, 50GB storage
Recommended: GPU with 16GB+ VRAM (NVIDIA RTX 3090/4090 or A100), 32GB+ RAM, 100GB+ SSD

7.2 Software

• Python 3.8+

• PyTorch 2.0+

• CUDA 11.8+ (for GPU training)

• timm (Vision Transformers)

• albumentations (Augmentation)

• sklearn (Metrics)

8. Conclusion

This training pipeline represents a state-of-the-art approach to GI endoscopy image classification, combining advanced Vision Transformer architectures, modern training techniques, and memory-efficient optimizations. The system is designed for production deployment with TorchScript optimization, FastAPI backend, Grad-CAM explainability, and RESTful API interface.

Key Takeaways:

• High Resolution Matters: 384px captures important medical details

• Ensemble Improves Reliability: Combining models reduces errors

• Advanced Augmentation: MixUp and TTA significantly improve performance

• Class Imbalance Handling: Focal Loss is crucial for medical datasets

• Memory Optimization: Enables training on consumer GPUs

Report Generated: November 11, 2025