GI Endoscopy AI Diagnostic Portal

Complete Project Report

Project: Advanced AI-Powered Gastrointestinal Endoscopy Diagnostic System

Date: November 11, 2025

Version: 1.0

Status: Production Ready

Executive Summary

The GI Endoscopy AI Diagnostic Portal is a complete end-to-end medical AI system that combines state-of-the-art Vision Transformer models with an intuitive web interface for automated classification of gastrointestinal endoscopy images. The system provides 23-class classification, explainable AI with Grad-CAM visualizations, ensemble predictions, and production-ready deployment.

Key Achievements

• Advanced ML Models: Ensemble of DeiT3 and ViT Base Vision Transformers

• High Accuracy: Optimized for medical image classification

• Explainability: Grad-CAM heatmaps for clinical interpretation

• Scalable Architecture: Docker-based deployment

• User-Friendly Interface: Modern React frontend with Tailwind CSS

• Production Ready: Complete CI/CD and deployment pipeline

Machine Learning Pipeline

Model Architecture

The system uses an ensemble of two Vision Transformer models: DeiT3 Small Patch16 384 (~22M parameters) and ViT Base Patch16 384 (~86M parameters). Both models are trained at 384×384 resolution with advanced data augmentation including MixUp, Focal Loss, Test-Time Augmentation, and Label Smoothing.

Parameter	Value
Image Size	384×384
Batch Size	2 (Effective: 16)
Learning Rate	1e-5
Epochs	25
Optimizer	AdamW
Loss Function	Focal Loss (γ=2.0)
Augmentation	MixUp, TTA, Label Smoothing

Backend API System

Architecture

FastAPI-based RESTful API with TorchScript-optimized models. Features include custom Grad-CAM for TorchScript compatibility, multi-layer visualization, attention rollout, lesion localization, and uncertainty estimation. Supports advanced image preprocessing and real-time inference.

Key Endpoints

• POST /predict - Main prediction with explainability

• GET /health - Health check with model status

• POST /generate-report - PDF explainability report

• POST /preprocess - Image preprocessing preview

Frontend Portal

Technology Stack

Modern React 18.2 application with Tailwind CSS 3.3. Features include drag-and-drop upload, real-time image preprocessing controls, model selection, advanced visualization options, and interactive Grad-CAM heatmaps with opacity and colormap controls.

Key Features

• Drag & Drop Image Upload

• Model Selection (Ensemble/DeiT3/ViT)

• Image Preprocessing Controls

• Advanced Visualization Options

• Interactive Heatmap Controls

• Comparative Views

• PDF Report Generation

Deployment Architecture

Docker Configuration

Complete Docker setup with multi-stage builds for optimization. Backend uses Python 3.10-slim with all dependencies. Frontend uses Node.js 18 for build and Nginx alpine for production serving. Docker Compose configuration for easy deployment.

Deployment Options

• Docker Compose (VPS/Cloud)

• Railway.app (Auto-deploy)

• Render.com (Free tier)

• AWS/Azure/GCP (Enterprise)

• Kubernetes (Scalable)

Technical Specifications

System Requirements: CPU: 4+ cores, RAM: 8GB+, GPU: Optional (8GB+ VRAM recommended)

Backend: Python 3.10+, PyTorch 2.0+, FastAPI 0.104+

Frontend: Node.js 18+, React 18.2+, Tailwind CSS 3.3+

Models: TorchScript format, 100-300 MB per model

Inference Time: 0.3-0.8s (GPU), 1-3s (CPU)

API Performance: < 1 second response time, 10-20 req/s throughput

Features & Capabilities

Core Features

✓ 23-Class Classification

✓ Ensemble Predictions

✓ Real-Time Inference

✓ Grad-CAM Explainability

✓ Multi-Layer Visualization

✓ Attention Rollout

✓ Lesion Localization

✓ Uncertainty Estimation

✓ Image Preprocessing

✓ PDF Report Generation

Conclusion

The GI Endoscopy AI Diagnostic Portal represents a complete, production-ready system combining advanced ML models, explainable AI, modern web interface, scalable backend, and production deployment. The system enables faster diagnosis, better accuracy, clinical trust through explainability, and seamless workflow integration.

Key Strengths

• High Accuracy: Ensemble approach with advanced techniques

• Explainability: Multiple visualization methods

• User-Friendly: Modern, responsive interface

• Production-Ready: Complete deployment pipeline

• Scalable: Docker-based architecture

• Maintainable: Clean code structure

Report Generated: November 11, 2025 | Project Status: Production Ready