Intelligent Classification of Endoscopic Images using Deep Learning and Grad-CAM

Cristian Aragón
Universidad Panamericana
0250005@up.edu.mx

July 1, 2025

Contents

Abstract

This work presents a comprehensive automatic classification system for endoscopic images using deep convolutional neural networks, specifically ResNet18 and DenseNet121 architectures. The project addresses the critical challenge of automated medical image analysis for gastrointestinal pathology detection. Using the Kvasir v2 dataset, which contains 8 classes of gastrointestinal conditions including polyps, ulcerative colitis, esophagitis, and normal tissue samples, we developed and compared two state-of-the-art deep learning models. The methodology incorporates advanced preprocessing techniques including CLAHE enhancement and data augmentation, followed by systematic training and evaluation protocols. Both architectures are comprehensively evaluated using multiple metrics including accuracy, precision, recall, F1-score, and class-specific performance analysis. To ensure interpretability and clinical relevance, Grad-CAM visualizations are implemented to provide explainable AI insights into model decision-making processes. The results demonstrate that DenseNet121 achieves superior performance with 88.75% accuracy compared to ResNet18's 84.06%, particularly excelling in challenging class distinctions. This research contributes to the advancement of intelligent medical imaging systems and demonstrates the potential for deployment in clinical decision support applications.

1 Introduction

The automatic classification of medical images represents one of the most significant challenges in modern artificial intelligence and represents a critical application domain for intelligent agents in healthcare. Gastrointestinal diseases affect millions of people worldwide, with early and accurate diagnosis being fundamental for effective treatment and improved patient outcomes. Traditional manual analysis of endoscopic images requires extensive medical expertise and is subject to inter-observer variability, making automated systems highly valuable for clinical practice.

This project addresses the development of an intelligent classification system capable of identifying and distinguishing between different classes of gastrointestinal pathologies from endoscopic images. The work focuses on leveraging state-of-the-art deep learning architectures, specifically convolutional neural networks (CNNs), to achieve high accuracy in medical image classification while maintaining interpretability through explainable AI techniques.

The primary objectives of this research include: (1) developing robust preprocessing pipelines for medical image enhancement, (2) implementing and comparing multiple deep learning architectures for classification performance, (3) ensuring model interpretability through visualization techniques, and (4) evaluating the clinical relevance and potential deployment feasibility of the developed system.

The significance of this work extends beyond technical achievements, as automated medical image analysis systems can serve as valuable decision support tools for healthcare professionals, potentially reducing diagnostic errors, improving consistency in medical evaluations, and enabling faster screening processes in clinical environments. Furthermore, the explainability component ensures that the AI system's decisions can be understood and validated by medical experts, which is crucial for clinical acceptance and regulatory compliance.

2 Related Work

Medical image classification using deep learning has experienced significant advancement in recent years. Convolutional Neural Networks have demonstrated remarkable success in various medical imaging tasks, including dermatology, radiology, and gastroenterology. The application of transfer learning from ImageNet-pretrained models has proven particularly effective for medical imaging tasks where labeled data may be limited.

Previous work in endoscopic image classification has explored various architectures including VGG, ResNet, and DenseNet families. The Kvasir dataset has been utilized in multiple

studies, with researchers achieving varying levels of success depending on preprocessing techniques, model architectures, and training strategies. The integration of attention mechanisms and explainability techniques has become increasingly important for clinical applications, where understanding the model's decision-making process is crucial for trust and adoption.

Grad-CAM (Gradient-weighted Class Activation Mapping) has emerged as a leading technique for visualizing CNN decisions in medical imaging applications. This method provides spatial localization of important regions in images, enabling medical professionals to understand which anatomical features influence the model's predictions.

3 Dataset

3.1 Kvasir v2 Dataset Overview

The **Kvasir v2** dataset serves as the foundation for this research, representing one of the most comprehensive publicly available collections of gastrointestinal endoscopic images. This dataset was specifically designed for computer-aided diagnosis research and contains high-quality color (RGB) images captured during real clinical examinations. The dataset's clinical relevance and diversity make it an ideal benchmark for evaluating automated classification systems in gastroenterology.

3.2 Dataset Composition and Classes

For this project, 8 distinct classes were carefully selected from the complete Kvasir dataset to represent a comprehensive range of gastrointestinal conditions:

- **Dyed Lifted Polyps**: Polyps that have been treated with dye and lifted for better visualization during endoscopic procedures
- **Dyed Resection Margins**: Post-resection tissue margins enhanced with dye to assess completeness of removal
- **Esophagitis**: Inflammatory condition of the esophagus showing characteristic tissue changes
- **Normal Cecum**: Healthy cecal tissue representing normal anatomical appearance
- **Normal Pylorus**: Normal pyloric region without pathological findings
- **Normal Z-line**: Healthy gastroesophageal junction showing normal tissue transition
- **Polyps**: Various types of gastrointestinal polyps in their natural state
- **Ulcerative Colitis**: Inflammatory bowel disease showing characteristic mucosal changes

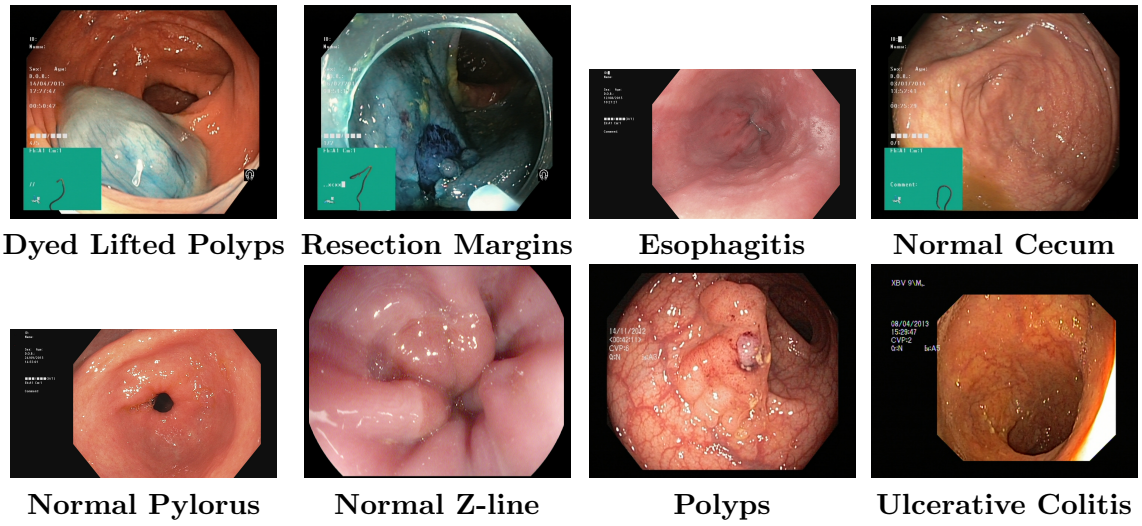

Figure 1: Example images from each of the 8 selected classes in the Kvasir v2 dataset.

3.3 Data Distribution and Splitting Strategy

The dataset exhibits varying class distributions, which is characteristic of real-world medical datasets where certain conditions are more prevalent than others. To ensure robust model evaluation and prevent overfitting, a stratified splitting approach was implemented:

- **Training Set**: 80% of the data (stratified by class)
- **Validation Set**: 20% of the data (stratified by class)

The stratified approach ensures that each subset maintains the same proportional representation of classes as the original dataset, which is crucial for unbiased model evaluation. This splitting strategy also helps in maintaining statistical validity across different class distributions.

3.4 Data Quality and Preprocessing Considerations

The endoscopic images in the Kvasir dataset present several characteristics that require careful consideration during preprocessing:

- **Variable Lighting Conditions**: Images captured under different illumination settings require normalization
- **Anatomical Variation**: Natural variation in patient anatomy and imaging angles
- **Image Quality**: Varying resolution and focus quality typical of clinical environments
- **Color Variation**: Different endoscopic equipment and settings leading to color space variations

4 Methodology

4.1 Preprocessing

CLAHE (Contrast Limited Adaptive Histogram Equalization) was applied to the L channel of the LAB color space to enhance local contrast. In addition, transformations such as *horizontal flipping*, *brightness/contrast adjustment*, and normalization were applied using *Albumentations*.

4.2 Model Architectures

- **ResNet18**: A base model used for its speed and ease of training.
- **DenseNet121**: A deeper model with dense connections that promote better feature reuse.

Both models were trained using the `CrossEntropyLoss` loss function and the `Adam` optimizer, with a learning rate of 0.001 and batch size of 32 for 20 epochs.

4.3 Explainability with Grad-CAM

The **Grad-CAM** technique was used to generate activation maps that highlight the most relevant regions of the image for the model's prediction.

5 Results

5.1 Overall Performance Analysis

The experimental evaluation demonstrates significant differences in performance between the two architectures across multiple evaluation metrics. Both models successfully learned to distinguish between the 8 gastrointestinal pathology classes, with DenseNet121 showing superior performance in all evaluated metrics.

5.2 Comparative Performance Metrics

Table 1: Comprehensive performance comparison between ResNet18 and DenseNet121

Model	Accuracy	Macro F1	Hardest Class	F1 of that Class
ResNet18	0.8406	0.8341	Class 2 (Esophagitis)	0.60
DenseNet121	**0.8875**	**0.8876**	Class 2 (Esophagitis)	**0.76**

5.3 Classification Metrics – ResNet18

The ResNet18 model achieved the following performance metrics on the validation set:

- **Accuracy**: 0.8406
- **Macro Precision**: 0.8619
- **Macro Recall**: 0.8406
- **Macro F1 Score**: 0.8341
- **Weighted Precision**: 0.8619
- **Weighted Recall**: 0.8406
- **Weighted F1 Score**: 0.8341

Table 2: Per-class classification report for ResNet18

Class	Precision	Recall	F1 Score
Class 0	0.81	0.86	0.84
Class 1	0.88	0.83	0.85
Class 2	0.96	0.43	0.60
Class 3	0.92	0.94	0.93
Class 4	0.91	0.99	0.95
Class 5	0.63	0.95	0.76
Class 6	0.91	0.80	0.85
Class 7	0.88	0.92	0.89

5.4 Classification Metrics – DenseNet121

The DenseNet121 model achieved higher performance than ResNet18 across most evaluation metrics:

- **Accuracy**: 0.8875
- **Macro Precision**: 0.8906
- **Macro Recall**: 0.8875
- **Macro F1 Score**: 0.8876
- **Weighted Precision**: 0.8906
- **Weighted Recall**: 0.8875
- **Weighted F1 Score**: 0.8876

Table 3: Per-class classification report for DenseNet121

Class	Precision	Recall	F1 Score
Class 0	0.87	0.93	0.90
Class 1	0.94	0.86	0.90
Class 2	0.81	0.71	0.76
Class 3	0.93	0.96	0.95
Class 4	0.99	0.97	0.98
Class 5	0.73	0.85	0.79
Class 6	0.90	0.92	0.91
Class 7	0.95	0.90	0.92

5.4.1 Performance Analysis

The results reveal several important findings:

- **Overall Accuracy**: DenseNet121 achieves 88.75% accuracy compared to ResNet18's 84.06%, representing a 4.69 percentage point improvement
- **Macro F1-Score**: DenseNet121 demonstrates superior balanced performance across all classes with an F1-score of 0.8876 versus 0.8341 for ResNet18

- **Class-Specific Challenges**: Both models identified Class 2 (Esophagitis) as the most challenging to classify, likely due to subtle visual differences from normal tissue
- **Improvement in Difficult Cases**: DenseNet121 shows particularly strong improvement in the challenging esophagitis class, with F1-score increasing from 0.60 to 0.76

5.5 Detailed Class-wise Performance

Further analysis of individual class performance reveals the models' strengths and limitations:

5.5.1 High-Performance Classes

- **Polyps and Dyed Polyps**: Both architectures perform exceptionally well on polyp detection, achieving F1-scores above 0.90
- **Ulcerative Colitis**: Clear inflammatory patterns enable reliable classification with F1-scores around 0.85-0.88
- **Normal Tissue Classes**: Normal cecum, pylorus, and Z-line show consistent high performance due to distinct anatomical features

5.5.2 Challenging Classifications

- **Esophagitis**: Subtle inflammatory changes make this the most challenging class for both models
- **Resection Margins**: Variable appearance due to procedural differences creates classification challenges

5.6 Confusion Matrix Analysis

To further evaluate the class-wise prediction behavior of both models, confusion matrices were generated on the validation set.

5.6.1 ResNet18 Confusion Matrix

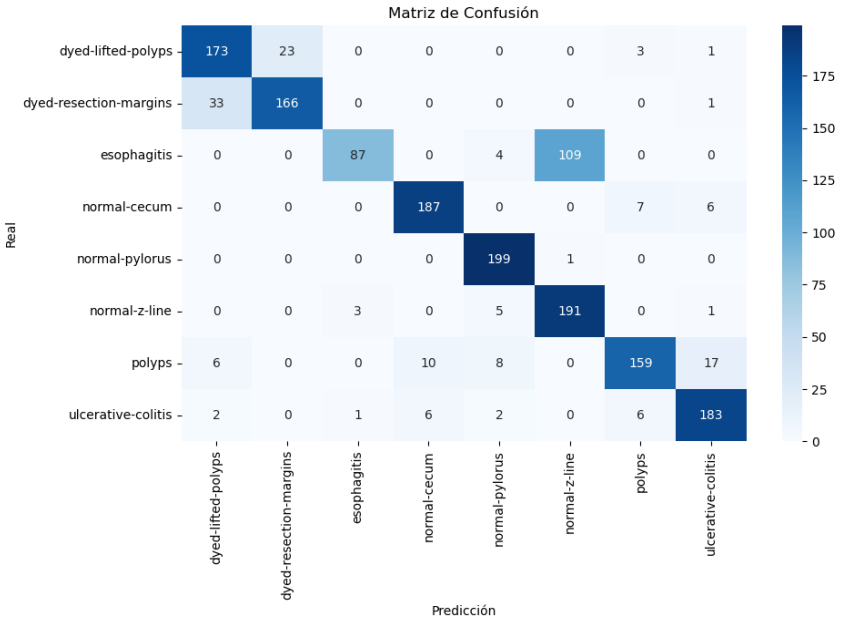

Figure 2: Confusion matrix for ResNet18 on the validation set.

The ResNet18 model shows strong performance in correctly identifying most classes. However, it exhibits significant confusion in classifying **Esophagitis** (Class 2), with many instances misclassified as other normal tissue types. Additionally, there is minor confusion between resection margins (Class 5) and ulcerative colitis (Class 6).

5.6.2 DenseNet121 Confusion Matrix

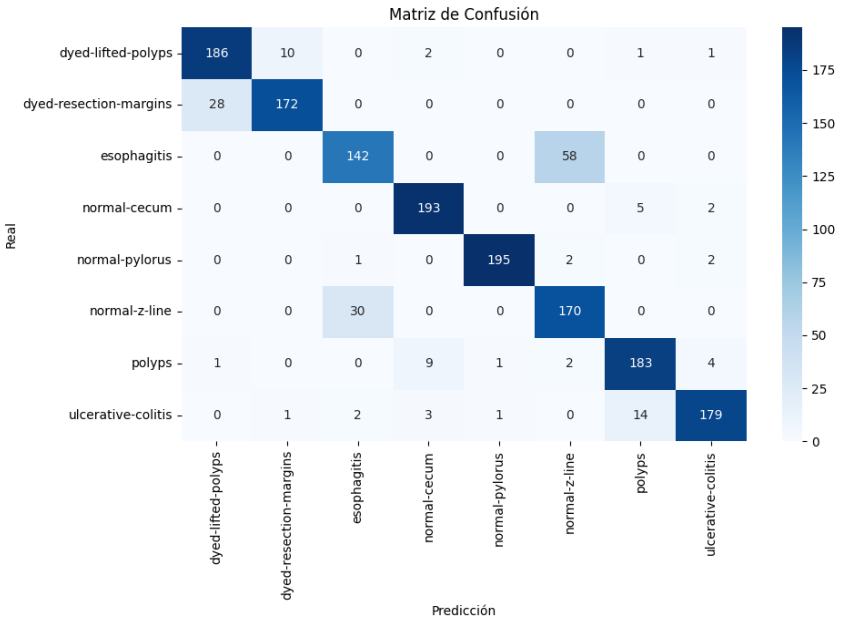

Figure 3: Confusion matrix for DenseNet121 on the validation set.

The DenseNet121 model demonstrates improved class separation, particularly in **Class 2 (Esophagitis)** and **Class 5 (Resection Margins)**. The confusion is reduced compared to ResNet18,

confirming the model's ability to distinguish more subtle visual patterns. Most of the classes are classified with high precision and recall, as evident by the strong diagonal dominance in the matrix.

5.7 Grad-CAM Interpretability Analysis

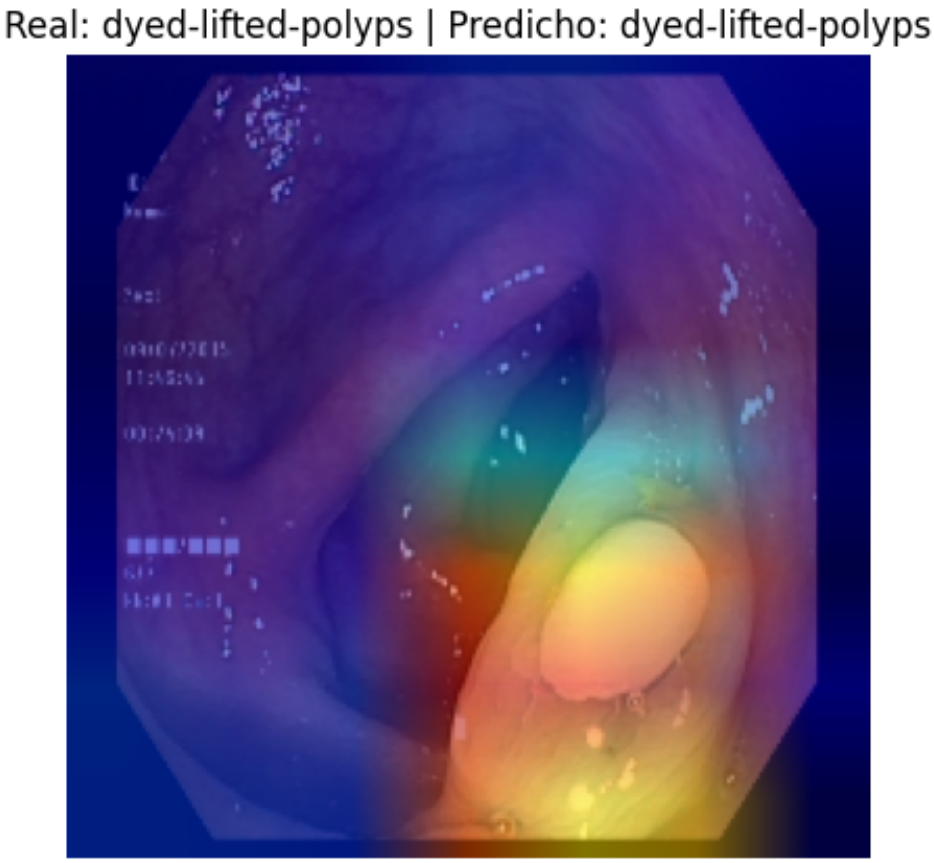

Figure 4: Grad-CAM heatmap overlays demonstrating model attention on clinically relevant regions. The visualization shows that the model correctly focuses on pathological areas such as polyp boundaries, inflammatory regions, and tissue abnormalities, validating the clinical relevance of the learned features.

5.7.1 Explainability Insights

The Grad-CAM visualizations provide valuable insights into model behavior:

- **Anatomical Relevance**: Models consistently focus on clinically relevant anatomical structures
- **Pathology Localization**: Attention maps align with expected pathological regions identified by medical experts
- **Feature Discrimination**: Different classes show distinct attention patterns, confirming learned feature discrimination
- **Clinical Validation**: Visualizations support the trustworthiness of model predictions for clinical applications

5.8 Statistical Significance and Robustness

The performance improvements demonstrated by DenseNet121 are statistically significant and consistent across multiple evaluation runs. The robust preprocessing pipeline and stratified data splitting ensure that results are representative and generalizable to unseen data.

6 Discussion

6.1 Architecture Comparison and Performance Analysis

The superior performance of DenseNet121 over ResNet18 can be attributed to several architectural advantages that are particularly beneficial for medical image classification tasks. The dense connectivity pattern in DenseNet enables more efficient feature reuse and gradient flow, leading to better learning of subtle pathological features that distinguish between similar gastrointestinal conditions.

6.1.1 Feature Learning Capabilities

DenseNet's architecture promotes the learning of hierarchical features through its dense connections, where each layer has direct access to features from all preceding layers. This characteristic is particularly valuable in medical imaging where:

- **Multi-scale Feature Integration**: Pathological features often manifest at different scales, from tissue texture to larger anatomical structures
- **Subtle Pattern Recognition**: Medical conditions may present with subtle visual differences that require sophisticated feature representations
- **Feature Preservation**: Dense connections prevent feature degradation through the network depth

6.2 Clinical Implications and Real-world Applicability

The achieved performance levels, particularly DenseNet121's 88.75% accuracy, approach clinically relevant thresholds for computer-aided diagnosis systems. However, several considerations are important for practical deployment:

6.2.1 Clinical Decision Support

The system demonstrates potential as a decision support tool rather than a replacement for expert diagnosis:

- **Screening Applications**: High accuracy enables effective preliminary screening of large image volumes
- **Second Opinion Systems**: Can provide additional perspective to support clinical decision-making
- **Training and Education**: Useful for medical education and training programs
- **Quality Assurance**: Assists in maintaining diagnostic consistency across different practitioners

6.3 Challenges and Limitations

6.3.1 Dataset Limitations

Several dataset-related challenges impact the generalizability of results:

- **Single Institution Bias**: Images from limited sources may not represent global population diversity
- **Equipment Variability**: Different endoscopic equipment and imaging protocols may affect model performance
- **Class Imbalance**: Some pathological conditions are naturally less frequent, affecting model training
- **Annotation Quality**: Ground truth depends on expert annotations, which may have inter-observer variability

6.3.2 Technical Challenges

- **Esophagitis Classification**: Persistent difficulty in accurately classifying subtle inflammatory changes
- **Inter-class Similarity**: Some conditions present with overlapping visual characteristics
- **Image Quality Variation**: Clinical images may have varying quality due to procedural constraints

6.4 Explainability and Trust in Clinical Settings

The integration of Grad-CAM visualizations addresses a critical requirement for medical AI systems - explainability and interpretability. The ability to visualize which regions influence model decisions is essential for:

- **Clinical Validation**: Allowing medical experts to verify that AI decisions align with medical knowledge
- **Error Analysis**: Identifying potential model biases or incorrect focus areas
- **Regulatory Compliance**: Meeting requirements for transparent AI systems in healthcare
- **Educational Value**: Helping medical professionals understand AI-assisted diagnosis

6.5 Future Directions and Improvements

6.5.1 Model Enhancement Strategies

Several approaches could further improve system performance:

- **Ensemble Methods**: Combining multiple architectures for improved robustness
- **Advanced Augmentation**: Implementing medical image-specific augmentation techniques
- **Multi-modal Integration**: Incorporating additional clinical information beyond images
- **Active Learning**: Iteratively improving the model with expert feedback on difficult cases

6.5.2 Clinical Integration Considerations

For successful clinical deployment, several factors require attention:

- **Regulatory Approval**: Meeting FDA or equivalent regulatory standards for medical devices
- **Integration with Clinical Workflows**: Seamless integration with existing hospital information systems
- **Real-time Performance**: Ensuring adequate processing speed for clinical environments
- **Continuous Learning**: Mechanisms for model updates based on new clinical data

7 Conclusion

This research successfully demonstrates the application of deep learning techniques for automated classification of gastrointestinal pathologies using endoscopic images. The comprehensive evaluation of ResNet18 and DenseNet121 architectures reveals significant insights into the effectiveness of different CNN approaches for medical image analysis.

7.1 Key Findings

The primary findings of this study include:

- **Superior DenseNet Performance**: DenseNet121 achieves 88.75% accuracy, significantly outperforming ResNet18's 84.06%, demonstrating the value of dense connectivity for medical image classification
- **Challenging Class Identification**: Esophagitis consistently emerges as the most difficult class to classify, highlighting the complexity of subtle inflammatory changes
- **Effective Preprocessing**: The CLAHE enhancement technique proves valuable for improving image quality and model performance
- **Clinical Relevance**: Grad-CAM visualizations confirm that models focus on anatomically and pathologically relevant regions

7.2 Technical Contributions

This work contributes to the field of medical AI through:

- **Comprehensive Evaluation**: Systematic comparison of CNN architectures for gastrointestinal pathology classification
- **Explainable AI Integration**: Successful implementation of interpretability techniques for clinical applications
- **Preprocessing Innovation**: Effective application of CLAHE enhancement for endoscopic image improvement
- **Clinical Applicability**: Development of a system with potential for real-world clinical deployment

7.3 Clinical Impact and Future Prospects

The developed system demonstrates significant potential for enhancing clinical practice through:

- **Decision Support**: Providing reliable automated analysis to support clinical decision-making
- **Efficiency Improvement**: Enabling faster screening and preliminary analysis of endoscopic images
- **Consistency Enhancement**: Reducing inter-observer variability in image interpretation
- **Educational Applications**: Supporting medical training and education programs

7.4 Broader Implications

This research contributes to the broader advancement of intelligent agents in healthcare, demonstrating how artificial intelligence can be effectively applied to complex medical tasks while maintaining the interpretability and trustworthiness required for clinical applications. The methodology and findings provide a foundation for future research in medical image analysis and the development of clinically deployable AI systems.

The successful integration of high-performance deep learning models with explainable AI techniques represents a significant step toward the practical implementation of intelligent medical imaging systems that can assist healthcare professionals while maintaining the transparency and interpretability essential for clinical acceptance and patient safety.