Deep Learning-based Classification of Endoscopic Images with MobileNetV2 and Preprocessing

Valeria Andrade
Universidad Panamericana
Aguascalientes, México
valeria.andrade.monreal@gmail.com

***Abstract*—This paper presents a deep learning approach for classifying gastrointestinal endoscopic images into three disease categories: esophagitis, polyps, and ulcerative colitis. A dataset of 3,000 images was curated and split into training, validation, and test sets, with perceptual hashing applied to prevent duplicate leakage. The proposed pipeline combines advanced preprocessing to remove ScopeGuide artifacts, extensive data augmentation, and transfer learning using a MobileNetV2 backbone pretrained on ImageNet. Preliminary results indicate that this approach achieves promising classification accuracy, supporting the feasibility of automated diagnosis in clinical practice.**

I. Introduction

Gastrointestinal (GI) diseases represent a significant global health burden, with conditions such as esophagitis, polyps, and ulcerative colitis affecting millions of patients annually. Early and accurate diagnosis of these diseases is critical to improving treatment outcomes and reducing the risk of complications, including malignant transformation in the case of precancerous lesions.

The Kvasir dataset, a publicly available collection of endoscopic images, provides an important benchmark for developing automated classification systems for GI diseases. While several studies have utilized this dataset to explore CNN-based approaches, relatively few have addressed rigorous preprocessing tailored to domain-specific artifacts or systematically evaluated the impact of advanced augmentation strategies on model performance.

In this work, I present a deep learning pipeline designed to classify endoscopic images into three clinically relevant categories: esophagitis, polyps, and ulcerative colitis. Our approach integrates multiple components:

- A dedicated preprocessing routine to remove visual artifacts and standardize input images.
- An extensive data augmentation pipeline to increase robustness to variability in appearance.
- Transfer learning based on the MobileNetV2 architecture pretrained on ImageNet, followed by fine-tuning to adapt the model to the endoscopic domain.

Through this study, I aim to assess the feasibility and effectiveness of combining preprocessing and modern deep learning techniques for accurate, automated classification of GI endoscopic images, thereby contributing toward decision-support tools that may improve the quality and efficiency of gastroenterology practice.

II. Dataset Description

The dataset used in this work is the **Kvasir dataset**, which contains high-resolution gastrointestinal endoscopic images. For this experiment, we selected three classes:

- *Esophagitis*
- *Polyps*
- *Ulcerative colitis*

A total of 3,000 images were included, with 1,000 images per class. The images were split into training, validation, and test sets with proportions of 70%, 15%, and 15%, respectively.

To create consistent and clean splits while avoiding near-duplicate images appearing in multiple sets, we applied the following pipeline:

1. *Deduplication with perceptual hashing:*

   For every image, a perceptual hash (phash) was computed using the imagehash library.

   All images sharing the same hash were grouped together to ensure that visually similar duplicates (e.g., slightly cropped or compressed variants) would remain in the same split.

   This approach mitigates data leakage during evaluation.

2. *Randomized splitting of unique hashes:*

   The dataset was split into training, validation, and test subsets with proportions of 70%, 15%, and 15%, respectively.

   To ensure reproducibility, a fixed random seed (42) was used to shuffle hashes prior to allocation.

3. *Directory structure generation:*

For each split and class, target directories were created in a new folder named splited_dataset.

All files were then copied into their assigned locations.

4. *Final distribution per class:*

Each class contained approximately:

- 700 training images
- 150 validation images
- 150 test images

Due to the deduplication logic, some splits varied slightly if entire hash groups were assigned to one subset.

Overall, this preprocessing ensured a robust and leak-free dataset, preserving class balance and providing clean input data for model training and evaluation.

.

III. Method Details

1. Preprocessing

The images were preprocessed using a custom pipeline that:

- Removed the green ScopeGuide box present in many endoscopy images by detecting green regions in HSV color space and replacing them with the surrounding average color.
- Cropped the bright regions corresponding to the main field of view.
- Normalized the input resolution to 224×224 pixels.

This preprocessing was implemented as a Lambda transformation within the Albumentations augmentation pipeline.

2. *Data Augmentation*

Training images were augmented using the Albumentations library with the following transformations:

- Random 90-degree rotations and horizontal or vertical flips
- Shifts, scaling, and rotations up to ±15°
- Random brightness and contrast adjustments
- Gaussian blur, gamma correction, and hue-saturation-value shifts

Validation images were only preprocessed (green box removal and cropping) and resized to 224×224 pixels, without any augmentation.

3. Model Architecture

- *Chroma Features (12 dimensions):*
- We used **MobileNetV2**, pretrained on ImageNet, as the base feature extractor. The model architecture comprised:
- The frozen base model
- A global average pooling layer
- A dense layer with 256 ReLU units and L2 regularization (1e-4)
- A dropout layer (rate 0.3)
- A softmax classification layer with 3 outputs

After initial training with the base model frozen for 10 epochs, the entire model was unfrozen and fine-tuned for 20 additional epochs with a lower learning rate (1e-5).

4. Training Details

- Batch size: 32
- Optimizer: Adam
- Loss function: Categorical cross-entropy
- Callbacks:
  - Early stopping with patience of 3 epochs
  - TensorBoard logging
  - WandB tracking

Training was conducted in two phases:

- Feature extraction (frozen backbone)
- Fine-tuning (all layers trainable)

IV. Results

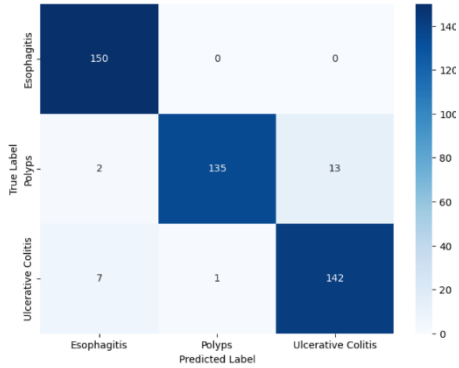
1. Overall Metrics:

- Accuracy: 94.89%
- Precision (macro): 94.39%
- Recall (macro): 93.56%
- F1 Score (macro): 93.47%
- Test Loss: 0.2026

2. Per class Metrics:

Class	Precision	Recall	F1 Score	Specificity
Esophagitis	94.34%	100%	97.09%	97.00%
Polyps	99.26%	90.00%	94.41%	99.67%
Ulcerative Colitis	91.61%	94.67%	93.11%	95.67%

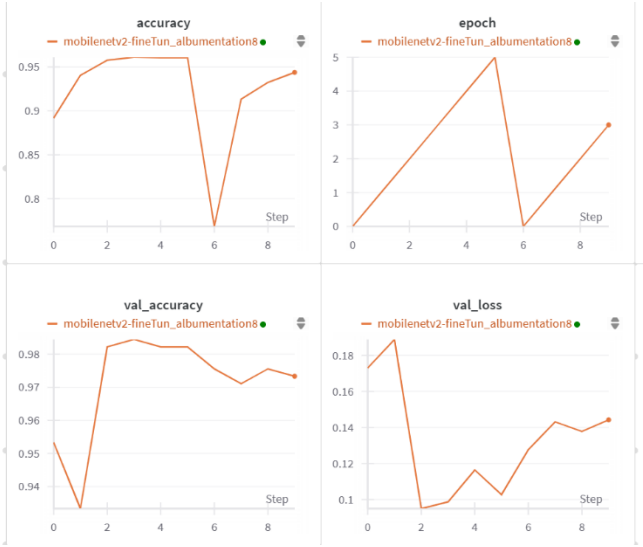
3. *Confusion Matrix:*



4. *Training Procedure and Convergence:*

The training process was divided into two stages. Initially, the pretrained MobileNetV2 convolutional base was frozen, and only the newly added classification head was trained. During this phase, the model quickly converged, reaching a validation accuracy of approximately 98% within three epochs. Early stopping restored the best weights from epoch 3 to prevent overfitting.

In the second stage, all layers of the backbone were unfrozen for fine-tuning with a reduced learning rate. As expected, the first epoch of fine-tuning showed a temporary drop in training accuracy (~68%) due to the re-adjustment of the pretrained weights. However, validation accuracy remained consistently high (>97%), demonstrating the model's robustness. Early stopping again selected the epoch with the lowest validation loss, ensuring that the final model retained optimal generalization performance.



V. Conclusion

This project successfully demonstrated the application of deep convolutional neural networks for the classification of endoscopic images into three clinically relevant categories: esophagitis, polyps, and ulcerative colitis. Using a MobileNetV2 backbone pretrained on ImageNet, combined with fine-tuning and extensive data augmentation, the model achieved strong performance across all metrics.

The final evaluation on the independent test set yielded an overall accuracy of 94.89%, with a macro-averaged precision of 94.39%, recall of 93.56%, and F1 score of 93.47%. Class-specific performance was also robust, with precision and recall values consistently above 90% for all categories. In particular, the recall for esophagitis reached 100%, while the model demonstrated high specificity across classes, indicating reliable discrimination between disease types.

The training procedure included a two-stage approach. Initially, the classifier head was trained with frozen convolutional layers, enabling rapid convergence and strong initial validation performance. Subsequently, all layers were unfrozen for fine-tuning with a lower learning rate. Although fine-tuning required only a few epochs due to early stopping, this phase further improved generalization without signs of overfitting.

Overall, the results confirm that transfer learning with a carefully designed augmentation pipeline and progressive training strategy can achieve high accuracy and robustness in medical image classification tasks. Future work could explore alternative architectures, such as EfficientNet or ensemble methods, and investigate the impact of larger datasets and more granular disease categories to further enhance model performance.

.