Gastrointestinal Image Classification

Christopher Oswaldo Márquez Reyes
Universidad Panamericana
Aguascalientes, México
0252751@up.edu.mx

Introduction

The automatic analysis of medical images using deep learning techniques has become a valuable tool to support healthcare professionals. In the field of gastrointestinal endoscopy in particular, accurate classification of pathologies enables early disease detection and significantly improves patient treatment outcomes.

In this project, an image classification model was developed using deep learning with the FastAI library and the pretrained ResNet50 model, adapted to the Kvasir v2 dataset. This dataset contains endoscopic images classified into multiple categories related to gastrointestinal medical conditions. To enhance the model's performance and generalization capability, advanced data augmentation techniques were applied using the Albumentations library.

The main objective was to build a robust model capable of distinguishing between different medical classes with high accuracy, evaluating metrics such as accuracy, precision, recall, and macro F1-score. This work is part of a broader effort to integrate artificial intelligence into clinical practice by providing tools that can assist in faster and more accurate diagnoses.

Dataset description

The dataset used in this work was Kvasir v2, a public dataset provided by the Simula Research Laboratory for Biomedical Computing and the Telemark Hospital in Norway. This dataset focuses on the analysis of endoscopic images from the gastrointestinal tract.
The dataset consists of 8 different visual classes representing common clinical conditions, including:

- Dyed-lifted-polyps
- Esophagitis
- Normal-cecum
- Normal-pylorus
- Normal-z-line
- Polyp
- Ulcerative-colitis
- Barrett's esophagus

Each class contains a variable number of color images with varying resolutions. The images were manually reorganized into training (`train`) and validation (`val`) folders to facilitate their use with the FastAI framework.

In this project, custom transformations were implemented using the Albumentations library, including normalization, rotation, contrast enhancement, and blur, in order to increase training data diversity and improve the model's generalization capability.

Method details

This project implements a complete pipeline for classifying endoscopic images using supervised deep learning techniques with the FastAI library and a pretrained ResNet50 model as the backbone. The following steps describe the methodology:

1. Environment Setup

The project was developed on Google Colab, which allowed for the use of free GPU resources. Google Drive was mounted to access the compressed dataset, which was extracted into the runtime environment.

2. Dataset Organization

The Kvasir v2 dataset was reorganized into two main folders: `train` and `val`, following the structure required by FastAI's `ImageDataLoaders.from_folder()`. Each class was placed in its corresponding subfolder within these main directories, allowing for automatic label assignment.

3. Data Augmentation with Albumentations

To improve the model's generalization and reduce overfitting risk, a custom data augmentation pipeline was defined using the Albumentations library. This pipeline included:

- CLAHE to improve local contrast
- RandomBrightnessContrast to simulate lighting variations
- GaussianBlur to simulate image blur
- HorizontalFlip and Rotate as geometric transformations
- Resize to 224x224 pixels and normalization using ImageNet statistics

These transformations were integrated into the FastAI pipeline through a custom `AlbumentationsTransform` class, enabling batch-level (`batch_tfms`) augmentations during training.

4. Data Loading

The `ImageDataLoaders.from_folder()` function was used with a batch size of 32, an initial resize to 256 px, and the custom batch transformations. This generated a `DataLoaders` object with a clear split between training and validation data.

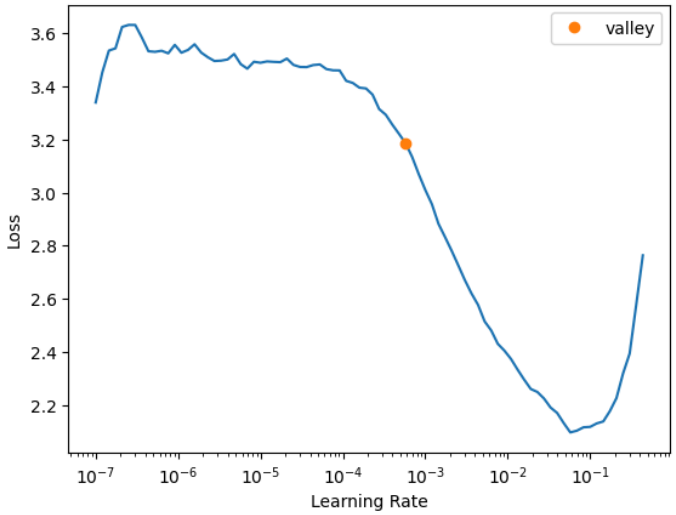
5. Model Definition and Metrics

A ResNet50 convolutional neural network pretrained on ImageNet was used for the classification task via FastAI's `vision_learner()` function.

In addition to standard accuracy, more representative metrics were used for imbalanced classes:

- Macro Precision
- Macro Recall
- Macro F1-score

These were implemented using `Precision`, `Recall`, and `F1Score` from the `fastai.metrics` module, with `average='macro'` to ensure equal weighting of all classes regardless of their frequency. This configuration is especially important in medical contexts, where misclassifying minority classes may have significant clinical implications.
Before training, the `lr_find()` method was run to identify an optimal learning rate by plotting loss against the learning rate on a logarithmic scale.



6. Model Training

The model was trained in two phases:

- Phase 1 (frozen): Initial training for 10 epochs with the ResNet50 convolutional base frozen, using a fixed learning rate of 1e-3.
- Phase 2 (fine-tuning): All layers were unfrozen, and the model was trained for an additional 20 epochs using a discriminative learning rate via `slice(1e-6, 1e-4)`.

In both phases, the `SaveModelCallback` was used to monitor validation loss (`valid_loss`) and automatically save the best model.

7. Model Evaluation

The final model was exported (.pkl) and evaluated using several metrics:

- Accuracy, Precision, Recall, and macro F1-score, using both FastAI and sklearn
- A confusion matrix was built to analyze class-wise performance
- The highest-loss predictions (top losses) from the validation set were visualized
- `multilabel_confusion_matrix` was used to extract detailed metrics per class: TP, TN, FP, FN

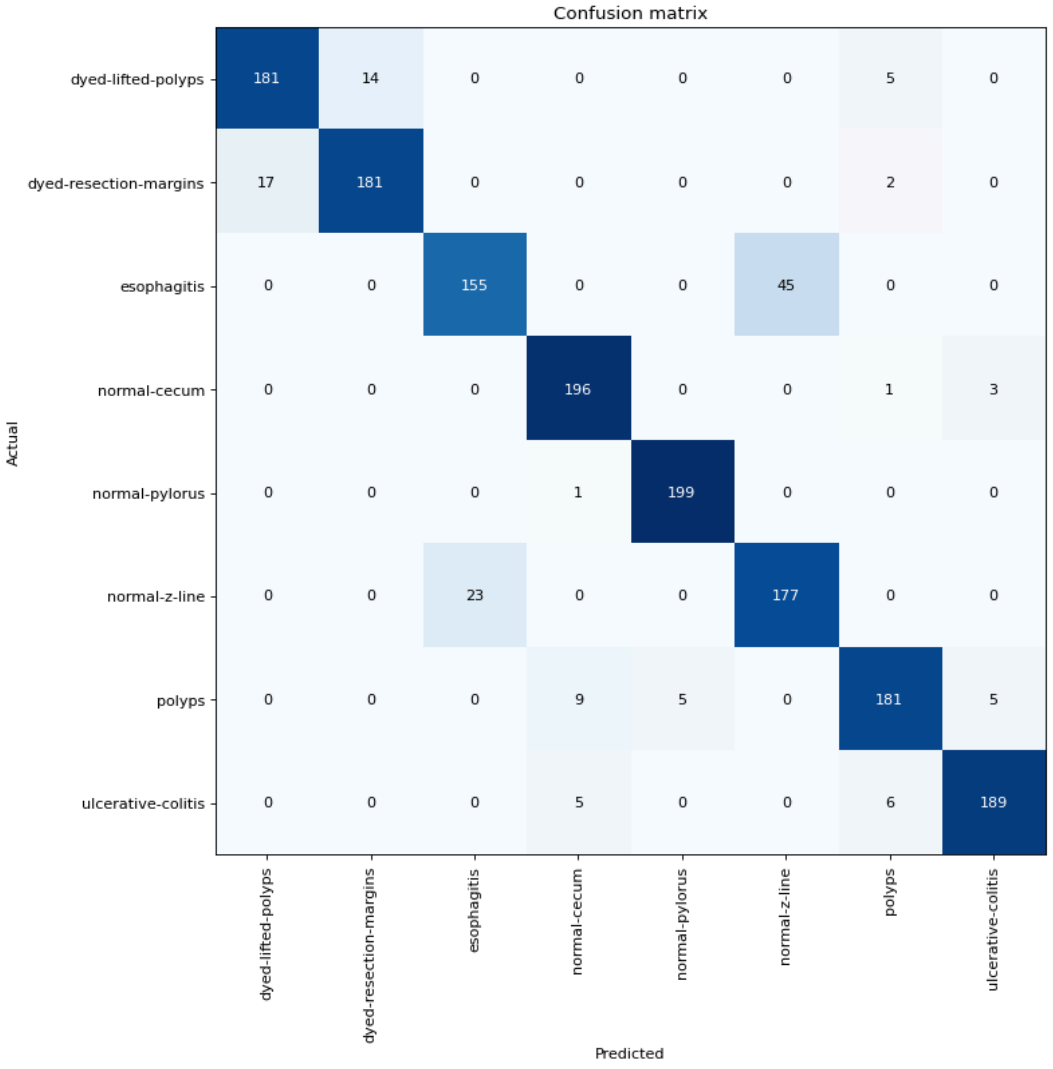
Results were compiled into a pandas DataFrame, exported as a `.csv` file, and presented in table format for analysis and interpretation.

Results

The model trained with the ResNet50 architecture and the Kvasir v2 dataset achieved strong overall performance, with a final macro F1-score of **0.9119** on the validation set. This metric reflects a good balance between precision and recall across all classes, regardless of sample imbalance.

🔹 **Confusion Matrix**
The confusion matrix (see Figure 1) shows that most classes were correctly classified, with high values along the diagonal. However, some noteworthy patterns of misclassification were observed:

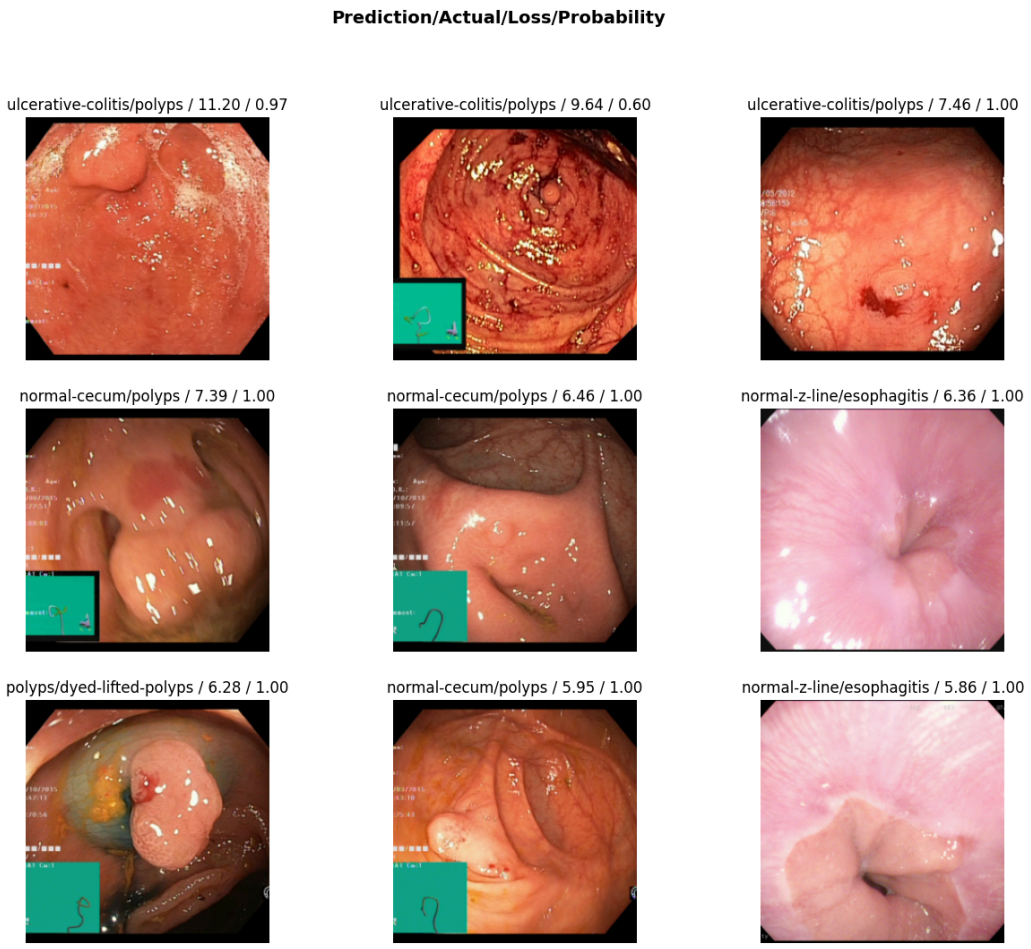
- *Esophagitis* was misclassified as *normal-z-line* 45 times, indicating visual similarities that could be addressed with more refined preprocessing or varied data.
- *Ulcerative-colitis* was misclassified as *polyp* 11 times, which is also evident in high-loss images due to their similar texture and coloration.
- *Normal-z-line* had a relatively low precision (0.7973) compared to other classes, mainly being confused with *esophagitis* (23 times).



🔹 Top Losses

Figure 2 presents the 9 most erroneous predictions (top losses), which reveal:

- Visually ambiguous cases where even experts may struggle to choose the correct diagnosis
- High confidence in incorrect predictions (probabilities close to 1.0), suggesting the model could benefit from techniques such as label smoothing or focal loss to reduce overconfidence.



🔹 Per-Class Metrics

A summary of the per-class metrics is provided, showing the performance balance and identifying specific areas of confusion.

class	precision	recall	f1_score	accuracy	TP	TN	FP	FN
dyed-lifted-polyps	0.9141	0.905	0.9095	0.9119	181	1383	17	19
dyed-resection-margins	0.9282	0.905	0.9165	0.9119	181	1386	14	19
esophagitis	0.8708	0.775	0.8201	0.9119	155	1377	23	45
normal-cecum	0.9289	0.98	0.9538	0.9119	196	1385	15	4
normal-pylorus	0.9755	0.995	0.9851	0.9119	199	1395	5	1
normal-z-line	0.7973	0.885	0.8389	0.9119	177	1355	45	23
polyps	0.9282	0.905	0.9165	0.9119	181	1386	14	19
ulcerative-colitis	0.9594	0.945	0.9521	0.9119	189	1392	8	11

Conclusion

This project demonstrated the effectiveness of deep learning models—specifically the ResNet50 architecture with the FastAI library—for multi-class classification of gastrointestinal endoscopic images using the Kvasir v2 dataset. The integration of advanced data augmentation techniques via Albumentations was crucial to improving the model's generalization, especially in the presence of class imbalance and visual similarities among clinical categories.

The model achieved a macro F1-score of **0.9119**, showing strong and consistent performance across most classes. However, areas such as *normal-z-line* and *esophagitis* showed higher confusion levels. These findings suggest potential improvements through the use of more specialized architectures, adaptive loss functions (e.g., focal loss), or expanding the dataset with more representative samples for problematic classes.

In conclusion, this system represents a step forward in developing AI-powered clinical support tools that can assist gastroenterology specialists in early detection and classification of diseases, contributing to faster, more objective, and more accurate diagnoses.