Classification of Histological Images of Colorectal Cancer Using Image Processing and Deep Learning

Luis Miguel Portugal Kegel
AI Engineering Student
Universidad Panamericana
Aguascalientes, México
luismi.porke@gmail.com

***Abstract*—This project presents a deep learning pipeline for classifying histology images from the Colorectal Histology MNIST dataset into eight tissue classes using a pre-trained ResNet50 model. The dataset, comprising 5,000 original images, is augmented to 24,000 training images through extensive offline data augmentation, followed by stratified splitting into training (24,000 images), validation (500 images), and test (500 images) sets. The model achieves a test accuracy of 94.80%, demonstrating robust classification performance with improved generalization due to the larger training set. This document details the dataset, methodology, code structure, results, and potential applications.**

I. Introduction

The Colorectal Histology MNIST dataset [1] contains 5000 original histological images, each 150x150 pixels in RGB format and stored as `.tif` files. Sourced from a 2016 study by Kather et al., the dataset is evenly distributed across eight classes: tumor, stroma, complex, lymphocytes, debris, mucosa, adipose, and empty, with 625 images per class. These hematoxylin and eosin (H&E) stained images capture microscopic tissue structures relevant to colorectal cancer analysis and are publicly available for research. The goal of this project is to develop a deep learning model to classify these images, aiding histopathological diagnostics by distinguishing tissue types based on their distinct visual features, such as dense nuclei in tumor regions versus clear backgrounds in empty regions.

To enhance model generalizability, the original 5,000 images are augmented to 24,000 training images using offline augmentations, including 0°–360° rotations, horizontal and vertical flips, brightness and contrast adjustments, Gaussian noise, and contrast limited adaptive histogram equalization (CLAHE). These augmented images are saved to disk, and the dataset is split into a stratified 80% training (24,000 images), 10% validation (500 images), and 10% test (500 images) sets to maintain class balance. Visual inspection of samples confirmed clear class distinctions, as illustrated in Figure 1. This pipeline leverages transfer learning with a pre-trained ResNet50 model, fine-tuned with a larger training set, offering potential applications in automated tissue analysis for clinical settings.

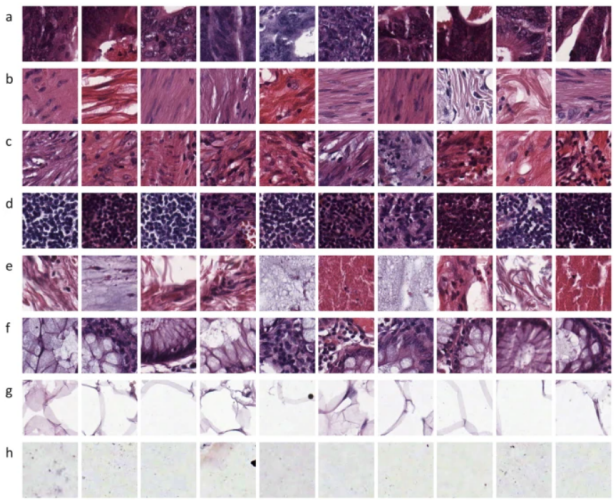

Fig. 1: Representative images from the dataset

II. Dataset Preparation

The dataset is prepared using the `preprocess.ipynb` script, which performs the following steps:

1) Collects image paths and labels from the dataset directory (`Kather_texture_2016_image_tiles_5000`), filtering for `.tif` files.
2) Applies offline data augmentation to the training set, generating 24,000 images (5 augmentations per original image) through 0°–360° rotations,

horizontal/vertical flips, brightness/contrast adjustments, Gaussian noise, and CLAHE, saving them to `Kather_texture_2016_augmented`.
3) Uses `sklearn.model_selection.train_test_split` to create stratified splits: 80% training (24,000 images), 10% validation (500 images), and 10% test (500 images).
4) Saves splits as CSV files (`train_split_augmented.csv`, `val_split.csv`, `test_split.csv`) in the `dataset_splits` directory.
5) Verifies class distribution to ensure each class retains approximately 12.5% of images per split post-augmentation.

Visual inspection confirmed data quality, with clear distinctions between classes and correct storage of augmented images.

III. Data Preprocessing and Augmentation

The `new_pytorch_dataset.py` script defines a custom PyTorch `Dataset` class (`HistologyDataset`) and configures `DataLoader` objects. Key features include:

- **Image Loading**: Images are loaded using OpenCV, converted from BGR to RGB.
- **Transformations**:
  - Training, Validation, and Test: Resizing to 224×224 and ImageNet normalization only, as augmentations are pre-applied and saved in `train_split_augmented.csv`.
- **DataLoaders**: Configured with a batch size of 32, multi-worker loading, and pin memory for efficiency, handling 24,000 training images (750 batches), 500 validation images (16 batches), and 500 test images (16 batches).
- **Error Handling**: Checks for missing files and invalid images.

The script verifies dataset sizes and batch properties, ensuring robust data handling without redundant augmentations.

IV. Model Training

The `new_resnet_training.ipynb` script trains a pre-trained ResNet50 model using transfer learning. Key steps include:

1) **Model Setup**: Loads ResNet50 with ImageNet weights, freezes all layers except `layer4` and the final fully connected layer (modified for 8 classes) to fine-tune with the larger dataset.
2) **Training Configuration**: Uses Adam optimizer (learning rate = 0.0001), CrossEntropyLoss, and trains for 10 epochs to account for the larger training set (24,000 images).
3) **Training Loop**: Tracks training and validation loss/accuracy, saving the model with the best validation accuracy.
4) **Evaluation**: Tests the best model on the test set, computing accuracy, precision, recall, and F1-score.
5) **Visualization**: Generates a confusion matrix and training history plots (loss and accuracy per epoch).

The increased training set size and fine-tuning adjustments contribute to improved performance over the baseline.

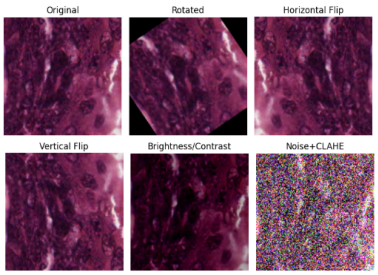

Fig. 2: training Samples

V. Results

The model achieves a test accuracy of 94.80%, with weighted precision, recall, and F1-score of 95.15%, 94.80%, and 94.88%, respectively. The best validation accuracy (91.8%) is reached at epoch 8. The confusion matrix (Figure 3) shows balanced performance across classes, with minor misclassifications. Training and validation loss/accuracy trends (Figure 4 and 5) indicate effective learning with no significant overfitting.

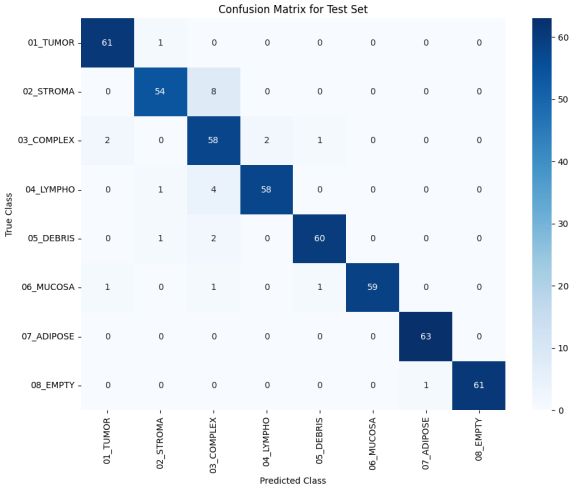

Fig. 3: Confusion Matrix for Test Set

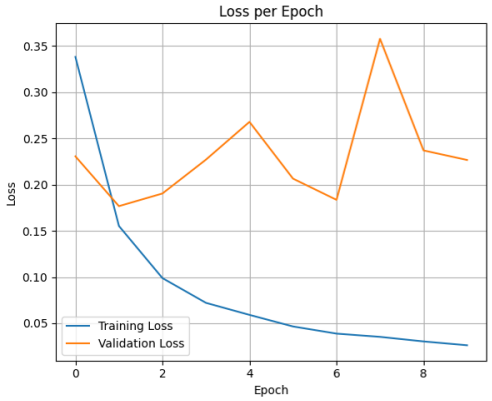

Fig. 4: Training and Validation Loss per Epoch

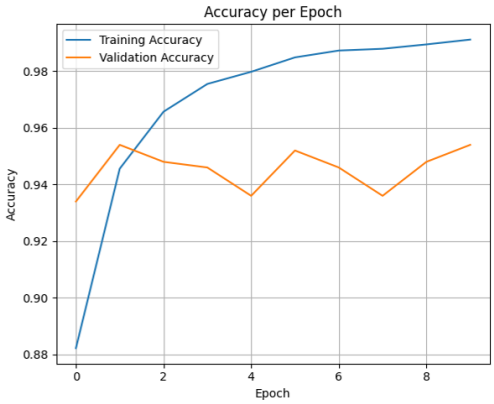

Fig. 5: Training and Validation Accuracy per Epoch

VI. Conclusion

This project demonstrates a robust pipeline for histology image classification using a pre-trained ResNet50 model. The high test accuracy and balanced metrics suggest effective transfer learning and data augmentation strategies. Future improvements could include unfreezing additional layers, experimenting with learning rate schedules, or testing alternative architectures like EfficientNet. The pipeline has potential applications in automated histopathological analysis, aiding medical diagnostics.

References

[1] Jakob Nikolas Kather. Colorectal histology mnist, 2016. Accessed: 2025-06-04.