Classification of Histological Images of Colorectal Cancer

Using Image Processing and Deep Learning

By Luis Miguel

Introduction

01

- **Colorectal cancer is a leading cause of global mortality.**

02

- **Histological image analysis is essential for diagnosis.**

Hematoxylin and eosin-stained tissues reveal cellular patterns critical for colorectal cancer diagnosis.

03

- **Manual classification is time-consuming and subjective.**

Classify eight tissue types: tumor, stroma, complex, lymphocytes, debris, mucosa, adipose, empty.

04

- **Goal: automate classification using machine learning.**

As AI engineering student, I aim to automate histological analysis to support pathologists and improve patient outcomes.

Objective & Motivation

1 Objective

- Automatically classify eight histological tissue types from the Colorectal Histology MNIST dataset

2 Motivation

- Reduce subjectivity and time in manual classification, supporting pathologists in colorectal cancer diagnosis.
- Improve on Kather et al.’s (2016) baseline accuracy of 0.874 through advanced preprocessing

DATASET

Colorectal Histology MNIST dataset.

1

5,000 original RGB images (150×150 px).

Starting set:

- Train: 4000 images
- Validation: 500
- Test: 500

2

8 classes

Tumor, Stroma, Complex, Lympho, Debris, Mucosa, Adipose, Empty.

3

Outcome

Generated 24,000 balanced training images (12.5% per class)
500 for validation.
500 for test.

DATASET

Colorectal Histology MNIST dataset.

tumor
stroma
complex
lymphocites
debris
mucosa
adipose
empty

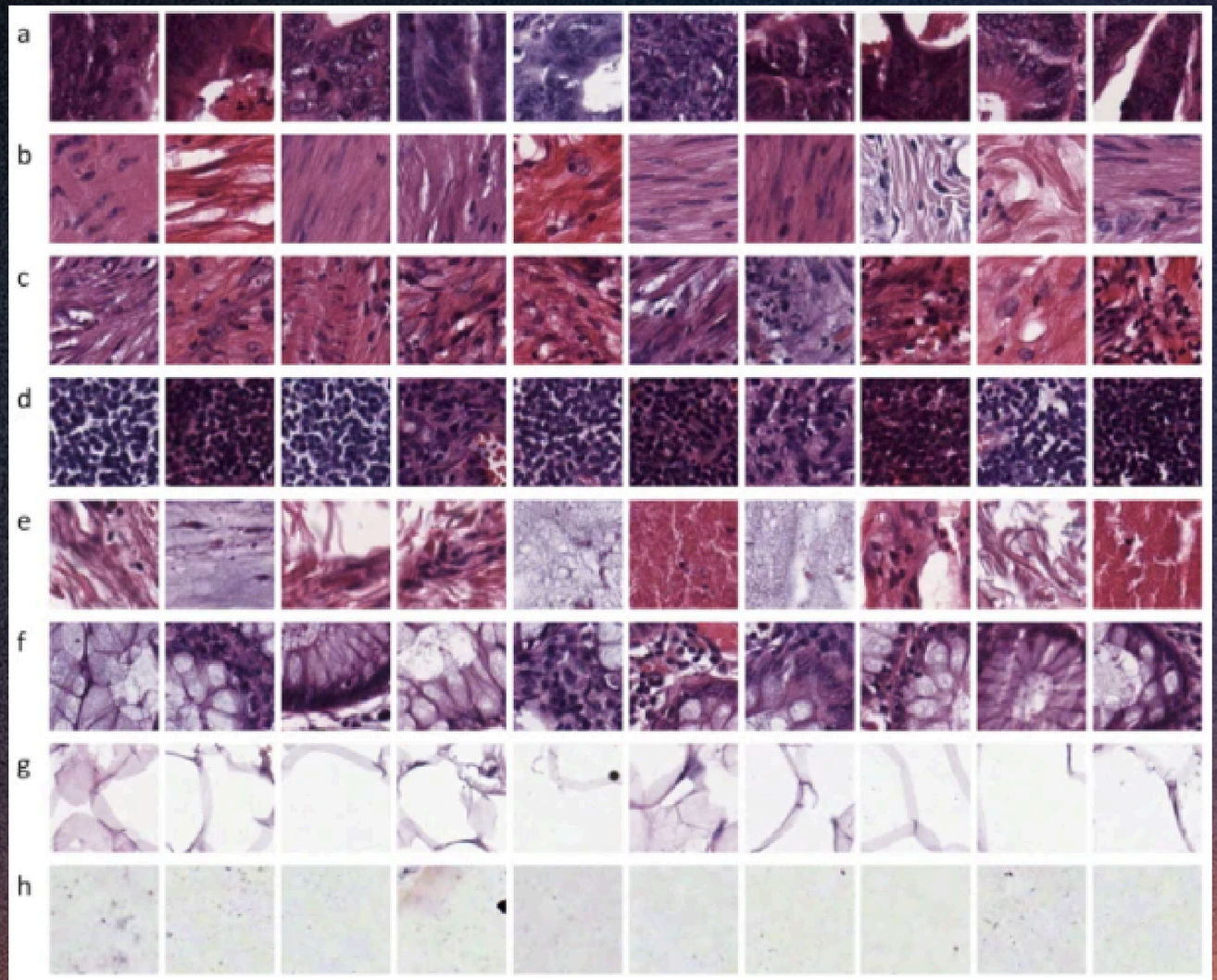


Fig. 1: Representative images from the dataset

Model Used

ResNet50

- Pre-trained on ImageNet
- 50-layer deep residual network
- Known for strong performance in image classification tasks.
- Fine-tuned on augmented dataset with 24,000 training images

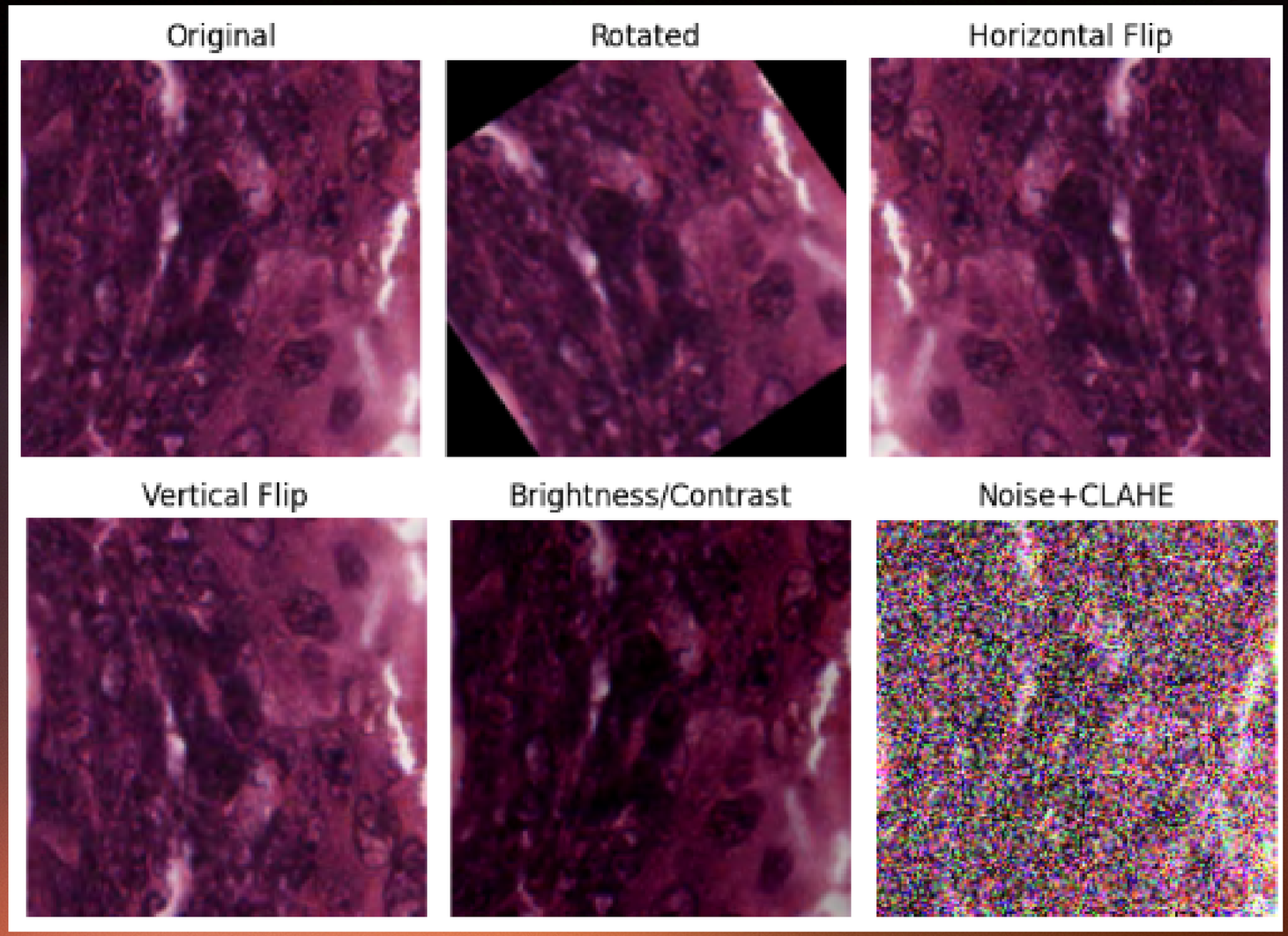
ResNet50

- Transfer Learning: Unfrozen layer4 + final FC layer
- Loss Function: CrossEntropyLoss
- Optimizer: Adam (LR = 0.0001)
- Epochs: 10
- Batch Size: 32
- class performance
- Early Stopping: Best model saved based on validation accuracy

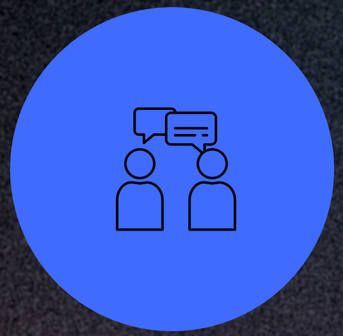
Data Augmentation

- 0°–360° rotations
- Horizontal and vertical flips
- Brightness and contrast adjustments
- Gaussian noise
- Contrast Limited Adaptive Histogram Equalization (CLAHE)
- Save augmented images to Kather_texture_2016_augmented.

Preprocessing Pipeline



Evaluation Metrics



Metrics:
Accuracy, Precision, Recall and F1-Score.

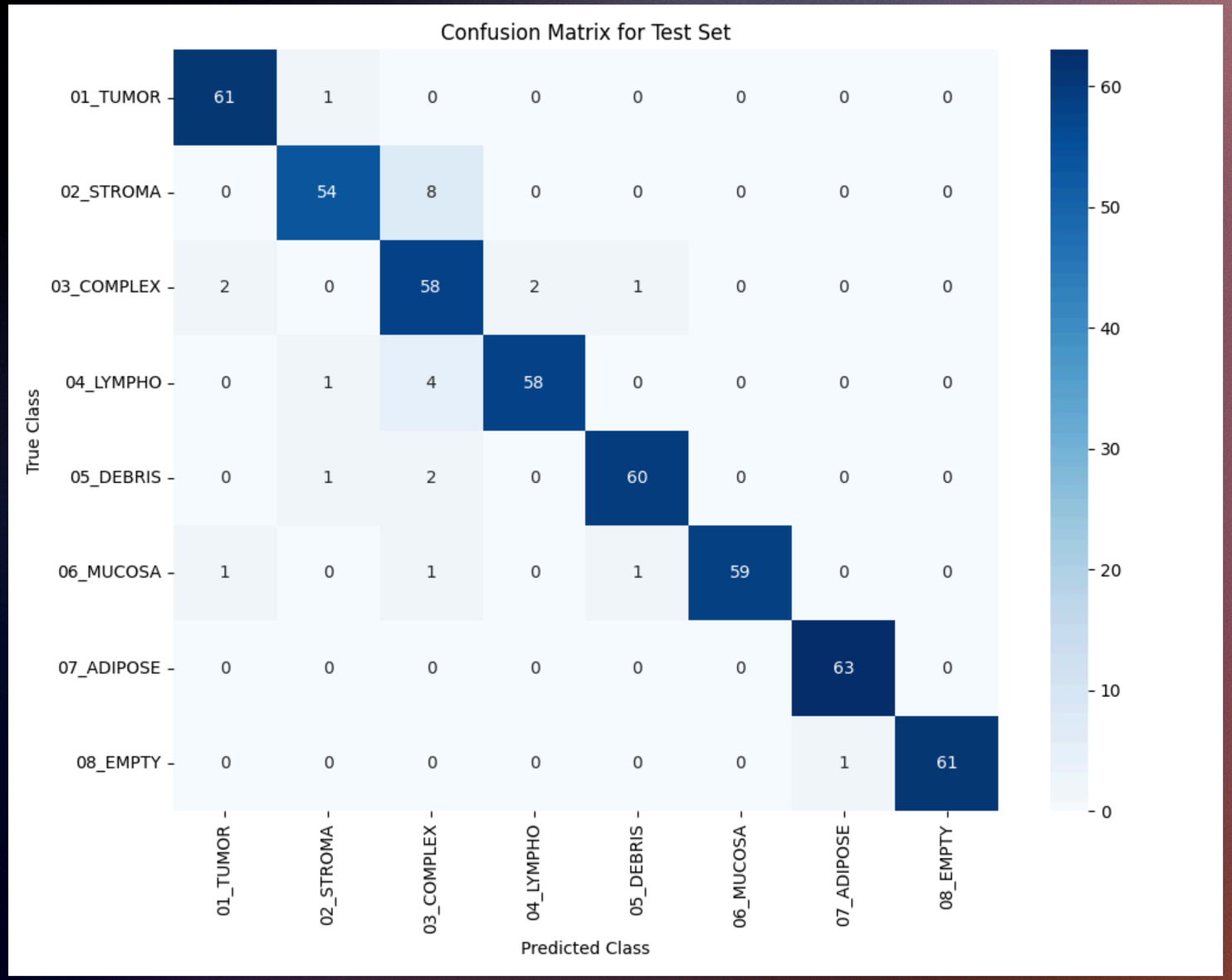
Confussion matrix and Loss/Accuracy trends for analysis

Results

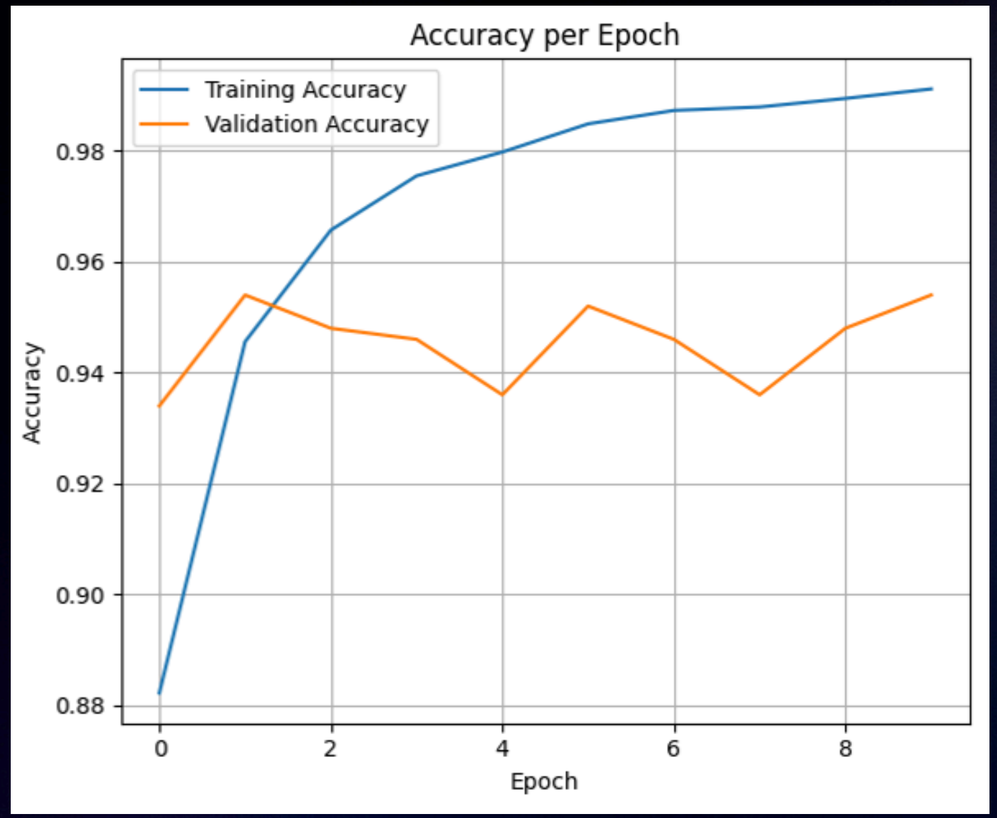
- Test Accuracy: 94.80%
- Weighted Precision: 95.15%
- Weighted Recall: 94.80%
- Weighted F1-Score: 94.88%
- Best Validation Accuracy: 91.8% (achieved at epoch 8)

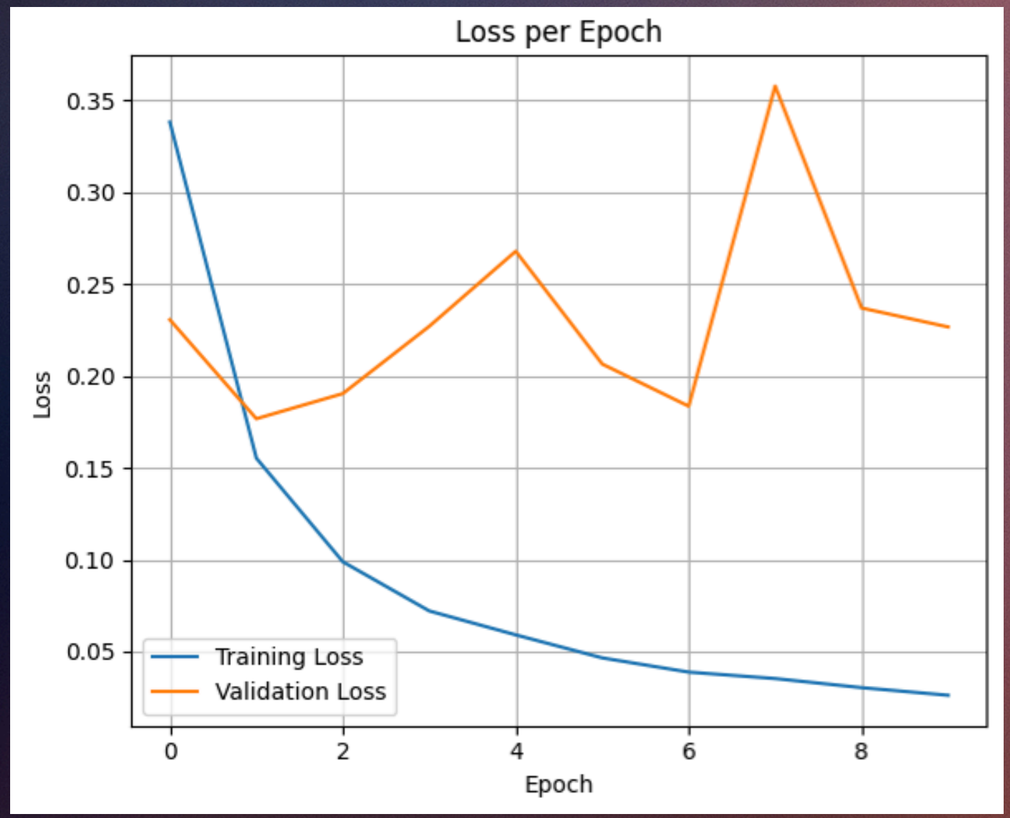
Visualizations

Normalized confusion matrix shows high accuracy for adipos, tumor and empty; While Stroma is the most confused.



Visualizations





Conclusion

- In this project, we successfully developed a deep learning pipeline for classifying histological images of colorectal cancer using a fine-tuned ResNet50 model. Through extensive data augmentation and transfer learning, the model achieved 94.80% test accuracy with balanced precision and recall across all eight tissue classes.

- This work highlights the potential of deep learning to assist in automated histopathological diagnostics, helping pathologists make faster and more accurate decisions.