Josiah Panak
https://github.com/PanakIllustrations/HW5-Kaggle-Competition

1. Overview

This project explores the use of deep learning for the automated detection of metastatic cancer in histopathology images. The task is formulated as a binary classification problem, where image patches are labeled as either containing tumor tissue or not.

A convolutional neural network (CNN) was trained on image patches extracted from whole-slide pathology scans. The resulting model achieved an accuracy of 96.6% and an area under the ROC curve (AUC) of 0.99, indicating high performance in distinguishing between tumor and normal tissue. These findings suggest that deep learning methods can support scalable and reliable analysis of pathology slides.

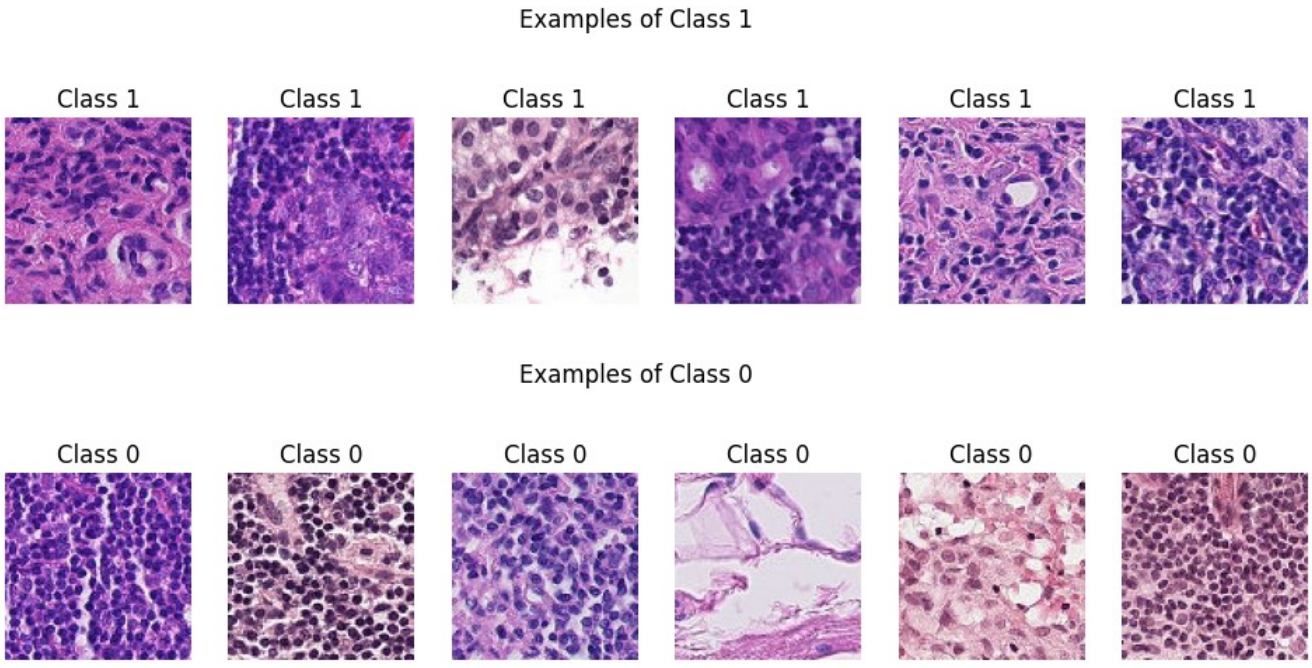
2. Data Summary

Dataset size: 220k labeled image patches

Class distribution: 59.5% non-tumor, 40.5% tumor

Data characteristics: 96 × 96px images with binary labels based on the center 32 × 32px area. The area outside of this center section is ignored when determining if a slide falls under a non-tumor classification.

Sample image analysis can reveal histological differences between classes, such as irregular cell morphology and disrupted tissue architecture in tumor samples. The main challenge of training a model on this data is the high visual complexity with low margin for error; slides are marked positive when containing even a single pixel of cancer.



Data Preprocessing

- pixel values are normalized from [0, 255] to [0, 1]

- Images are cropped to 32 × 32px area to focus on diagnostic regions

- Shuffle and split using train_test_split with stratification to ensure balanced classes.

Potential enhancement include data augmentation, stain normalization, and outlier removal to further improve model robustness.

3. Model Architecture

A custom CNN was design for histopathological image analysis:

- input size was 96 × 96 × 3

- layers: Multiple convolutional and pooling layers (32-129 filters), followed by dense layers

- Output: Sigmoid activation for binary classification

- Regularization: 50% dropout

This architecture results in ~2.3 million parameters and ~9MB model size, taking approximately 4 hours to train.

Training Strategy

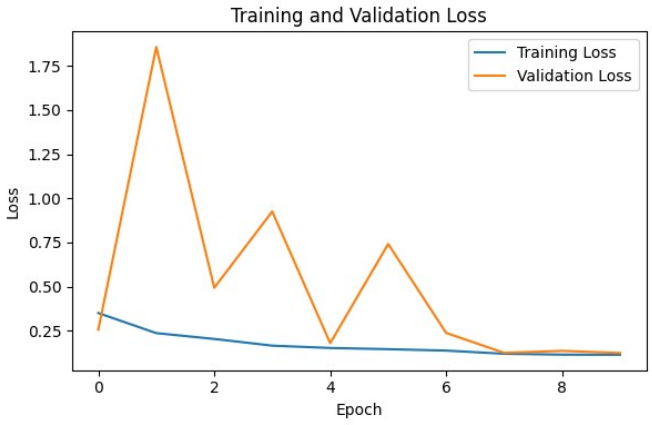
Optimizer: Adam with learning rate decay

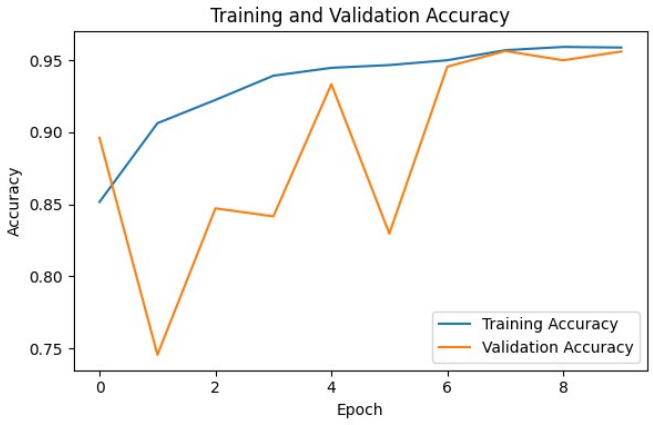
Loss Function: Binary cross-entropy

Batch Size: 32

Epochs: 10

```
Epoch 10/10
5501/5501 ━━━━━━━━━━━━━━━━━━━━ 1120s 204ms/step - accuracy: 0.9592 - auc: 0.9906 - loss: 0.1141 - val_accuracy: 0.9561
```





Training and validation accuracies remained closely aligned (~95.6%), indicating strong generalization.

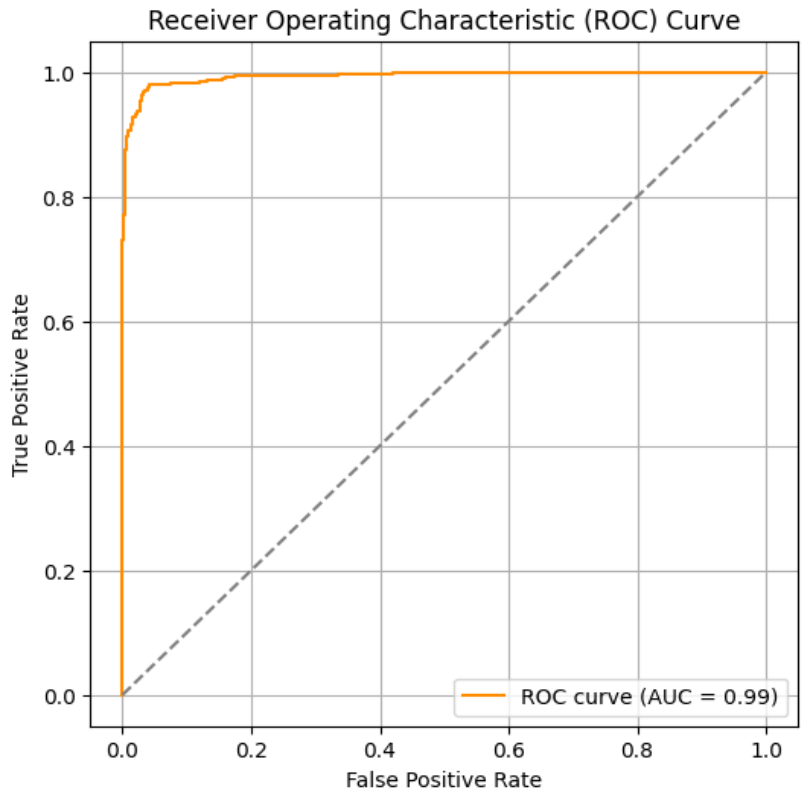
4. Results

Verify results with sample of 1000 images:

```
Accuracy: 0.966

Classification report:
              precision    recall  f1-score   support    Confusion matrix:
                                                          [[568  22]
           0       0.98      0.96      0.97       590     [ 12 398]]
           1       0.95      0.97      0.96       410

    accuracy                           0.97      1000
   macro avg       0.96      0.97      0.96      1000
weighted avg       0.97      0.97      0.97      1000
```



The AUC of 0.99 suggest excellent separability between benign and malignant cases.

Model Tuning

Used validation loss and accuracy to guide early stopping.

Batch size and learning rate were tuned for stability and speed.

Dropout layers were used to minimize over fitting.

5. Conclusion

The model demonstrates strong performance, achieving high accuracy and excellent AUC. While the results are promising, there remain opportunities for further improvement. Simple changes, such as applying data augmentation to reduce over-fitting or removing outliers that introduce noise, could have enhanced the model's robustness.