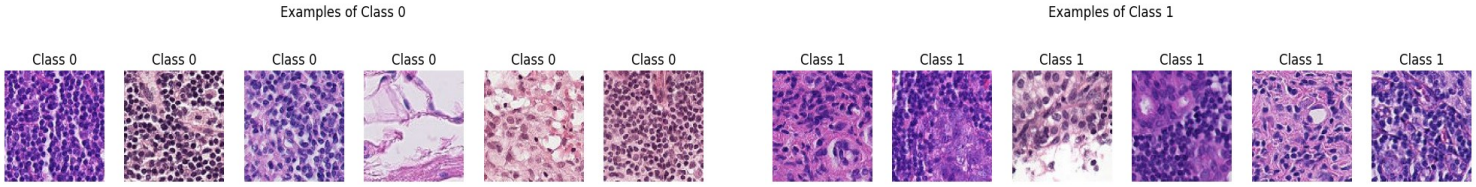
## Overview

This project seeks to create an algorithm to identify metastatic cancer in small image patches extracted from larger digital pathology scans. The detection of metastatic cancer in histopathology images can be classified as a binary image classification problem, as the training dataset of over 220k images has a binary label that is positive whenever there is at least one pixel of tumor tissue in the corresponding image.

## Data Analysis

*Class distribution:*
0 130908
1 89117
Percentage of tumor samples: 40.50%
Percentage of non-tumor samples: 59.50%

Random sample of six negative and positive images (regions outside the center box are ignored when determining the label):




*Data Cleaning:*
- Normalize pixel values ( converting integer between 0 and 255 to a float value between 0 and 1 )
- Cropping images to center 32 × 32px to remove potential false-negative errors from dataset
- shuffle and split using train_test_split with stratification to ensure balanced classes.

## Model Architecture:

The model is a Convolutional Neural Network (CNN) built using Keras. CNN capture spatial hierarchies through convolutional filters, making it especially effective at detecting features such as abnormal cell structures or densities that might indicate the presence of cancer.

*Simplified layer overview:*
Input: resized color image (e.g., 96×96×3)
Several Conv2D + MaxPooling layers for feature extraction
Dense layer(s) for classification
Output: Dense(1, activation='sigmoid') for binary prediction

## Results & Analysis:

*Training and Validation Metrics:*

Training and validation loss steadily decreased
Accuracy improved with each epoch and stabilized around 96%

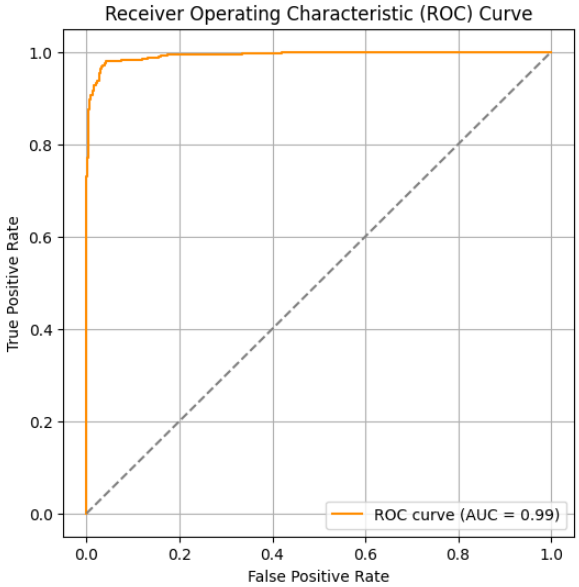
```
Epoch 1/10
5501/5501 ━━━━━━━━━━━━━━━━━━━━ 1179s 212ms/step - accuracy: 0.8172 - auc: 0.8850 - loss: 0.4267 - val_accuracy: 0.8961
Epoch 9/10
5501/5501 ━━━━━━━━━━━━━━━━━━━━ 1051s 191ms/step - accuracy: 0.9588 - auc: 0.9901 - loss: 0.1161 - val_accuracy: 0.9500
Epoch 10/10
5501/5501 ━━━━━━━━━━━━━━━━━━━━ 1120s 204ms/step - accuracy: 0.9592 - auc: 0.9906 - loss: 0.1141 - val_accuracy: 0.9561
```

*Performance metrics:*

```
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
```




## Conclusion

This model demonstrates strong performance, achieving high accuracy and AUC. While the results are promising, there were still some missed opportunities for further improvement, such as applying data augmentation to reduce over fitting or removing outliers that introduce noise.