**Kevin Qian**

## Problem Description and Data Overview

**Git:** https://github.com/KQian-lab/kagglehw5

**Problem:**

The task of this competition was to classify histopathological images of lymph node sections into either cancerous or non-cancerous categories.

### Data Overview

**Training labels:** Contains image IDs and the corresponding labels indicating the presence of cancer, 1, or absence, 0.

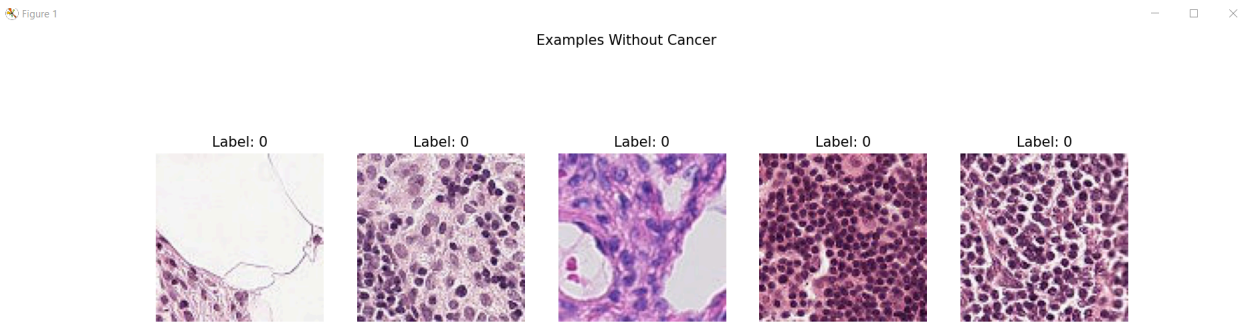
**Dimensions:** Images are processed into 96 x 96 pixels with three color channels.

## Data Size

- 220,025 Training Images
- 57,458 Test Images

## Exploratory Data Analysis and Data Cleaning

The data contained no missing values and image dimensions were ensured to remain consistent



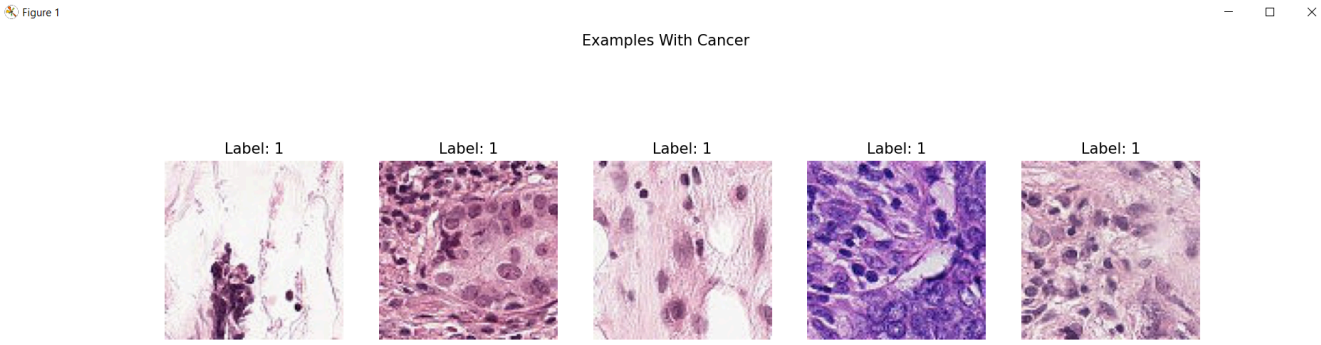
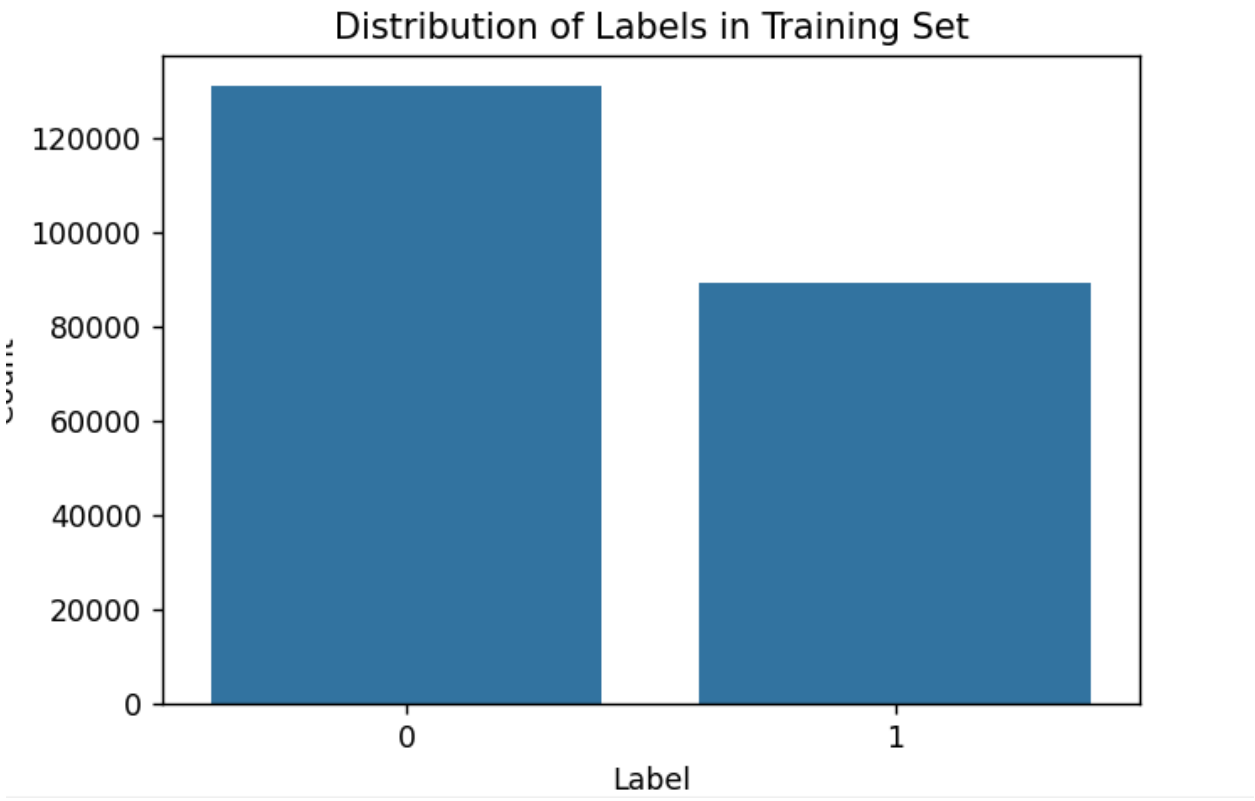


Figure 1
Examples With Cancer
Label: 1
Label: 1
Label: 1
Label: 1
Label: 1


Distribution of Labels in Training Set
120000
100000
80000
60000
40000
20000
0
0
1
Label

## Model Architecture

The model used is a Convolutional Neural Network for binary classification.

Convolutional Neural Networks are good at processing images because they can automatically detect patterns and features, such as edges and textures, which are important for classifying images. The output layer uses a sigmoid activation function which is better for binary classification tasks. In this case, identifying if an image is cancerous or not.

## Results and Analysis

Training and Validation:

- Epochs: 6
- Batch Size: 32
- Performance measurement: Accuracy and AUC

Issues Encountered:

- System Resources - Lack of GPU processing resulted in very long processing times for each epoch
    - This was resolved by using a much smaller random sample of images instead of the entire set

Hyperparameters:

- Learning Rate: Set by the Adam optimizer, 0.001
- Batch Size: 32

- Number of Filters in Convolutional Layers: 32, 64, 128
- Kernel Size: (3,3)
- Dropout Rate: 0.25 for convolutional layers, 0.5 in dense layers
- Number of Neurons in Dense Layer: 256
- Activation Functions: ReLU
- Epochs: 6

| Submission and Description | Private Score | Public Score |
|---|---|---|
| **submission1.csv**<br>Complete (after deadline) · 15s ago | 0.6883 | 0.7146 |

## Conclusion

This implemented CNN model is effective for processing histopathological images as cancerous or non-cancerous. I believe if it were possible to train the model on the entire dataset instead of a small random sample, the results would have been much better.