BrainStation Capstone Project

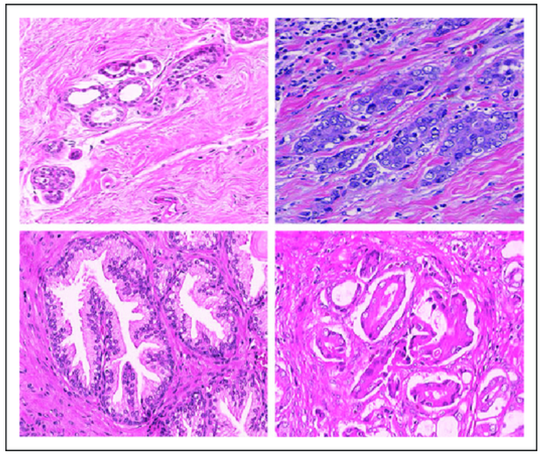
Sprint 3

Breast Tumor Classifier

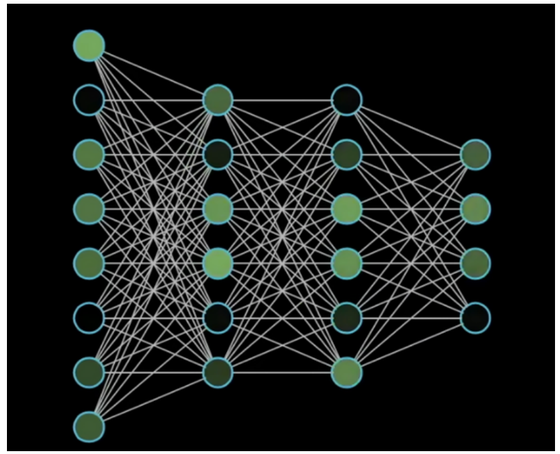
- Subject Area
- Problem Statement
- Project Vison
- Data Source
- Exploratory Data Analysis
- Data Preprocessing
- Computer Vision Models
- Model Metrics
- Model Evaluation
- Model Deployment
- Streamlit Demo App

Subject Area

Intersection of Medical Imaging, Machine Learning and Histopathology



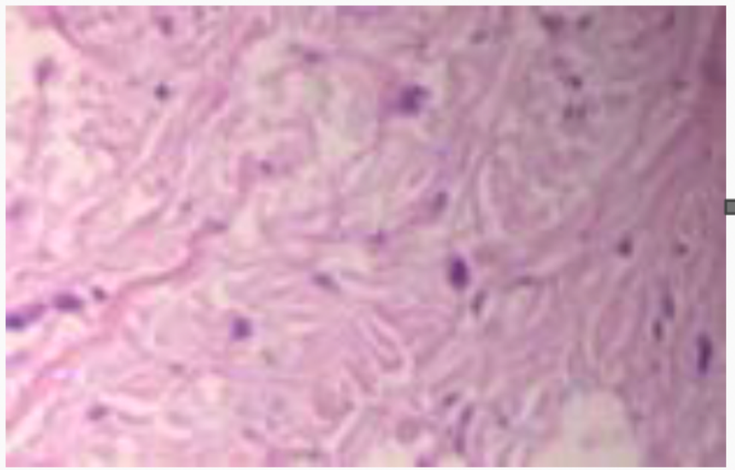
MEDICAL IMAGING



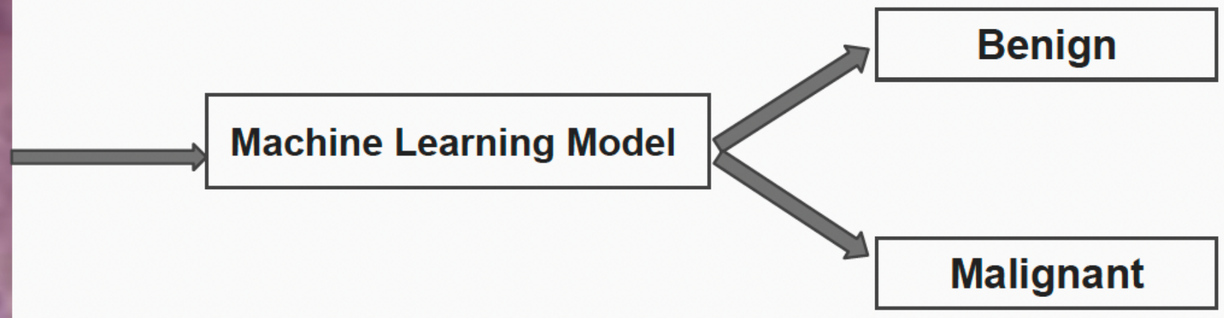
COMPUTER VISION

HISTOPATHOLOGY

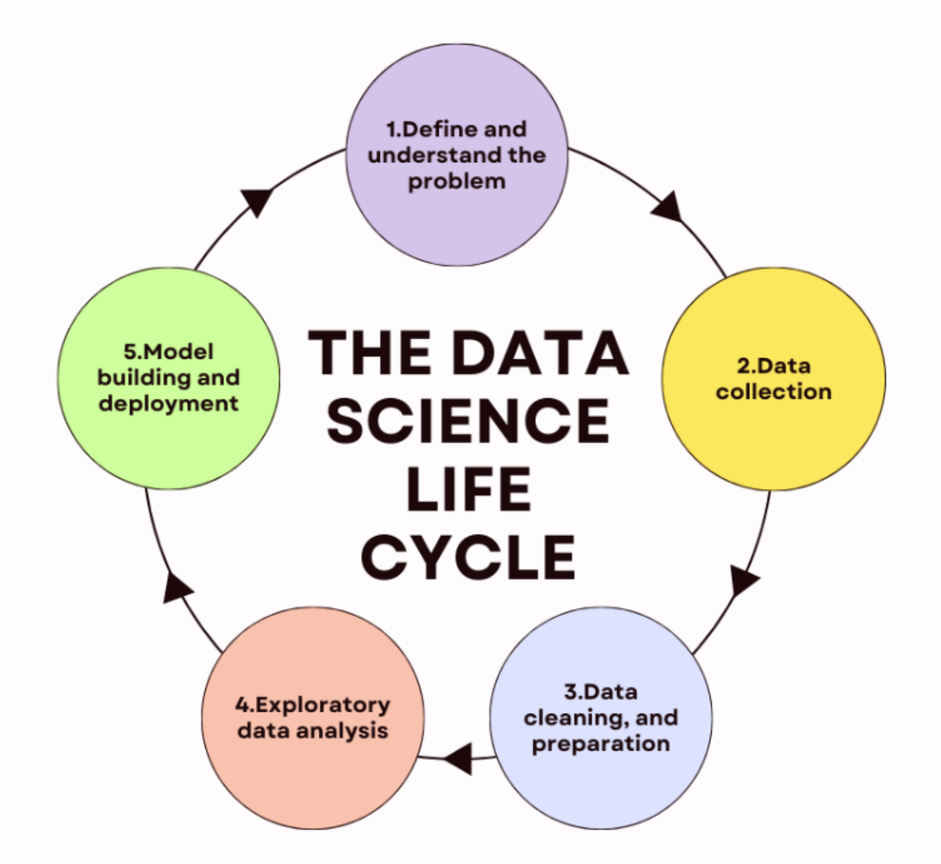
Problem Statement



Histopathological Image



Project Vision



Data Source

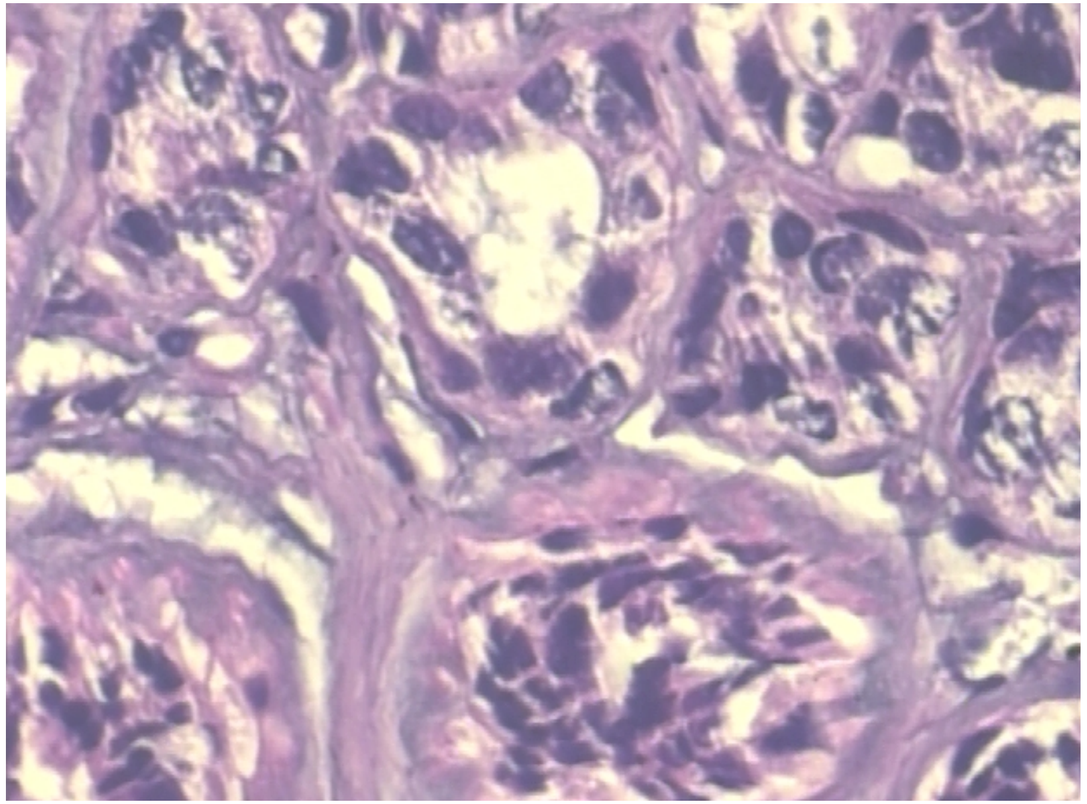
Breast Cancer Histopathological Database (BreakHis)

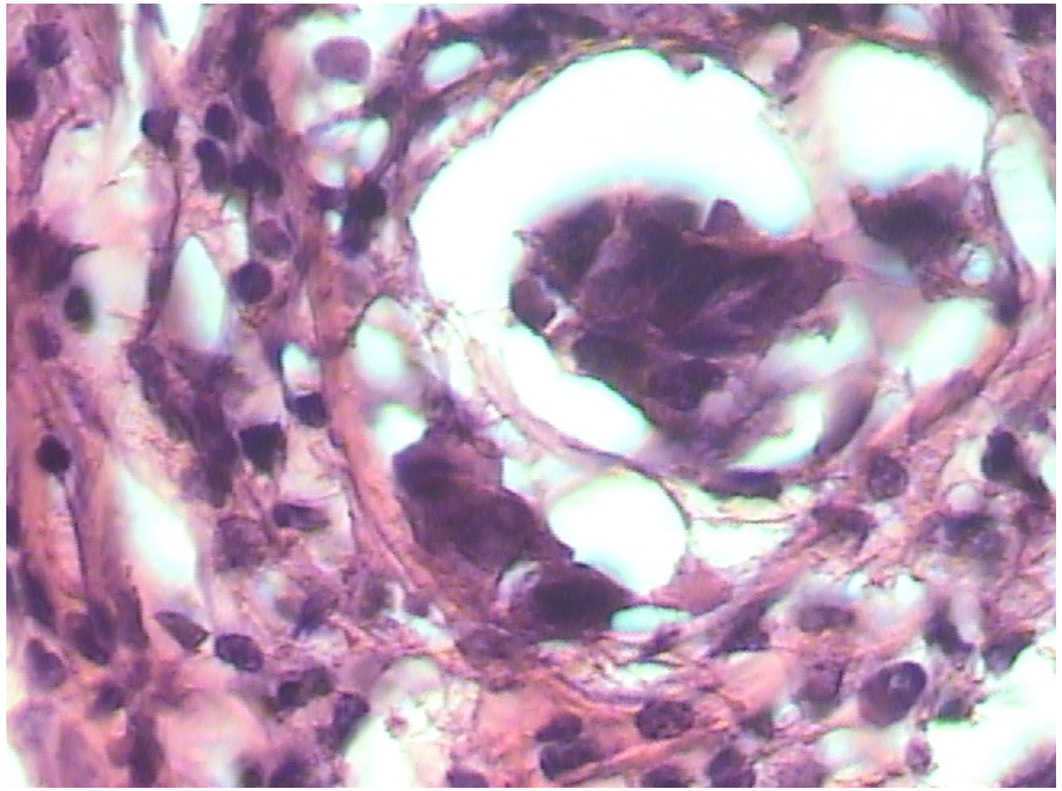
A Dataset for Breast Cancer Histopathological Image Classification

Fabio A. Spanhol *, Luiz S. Oliveira, Caroline Petitjean, and Laurent Heutte

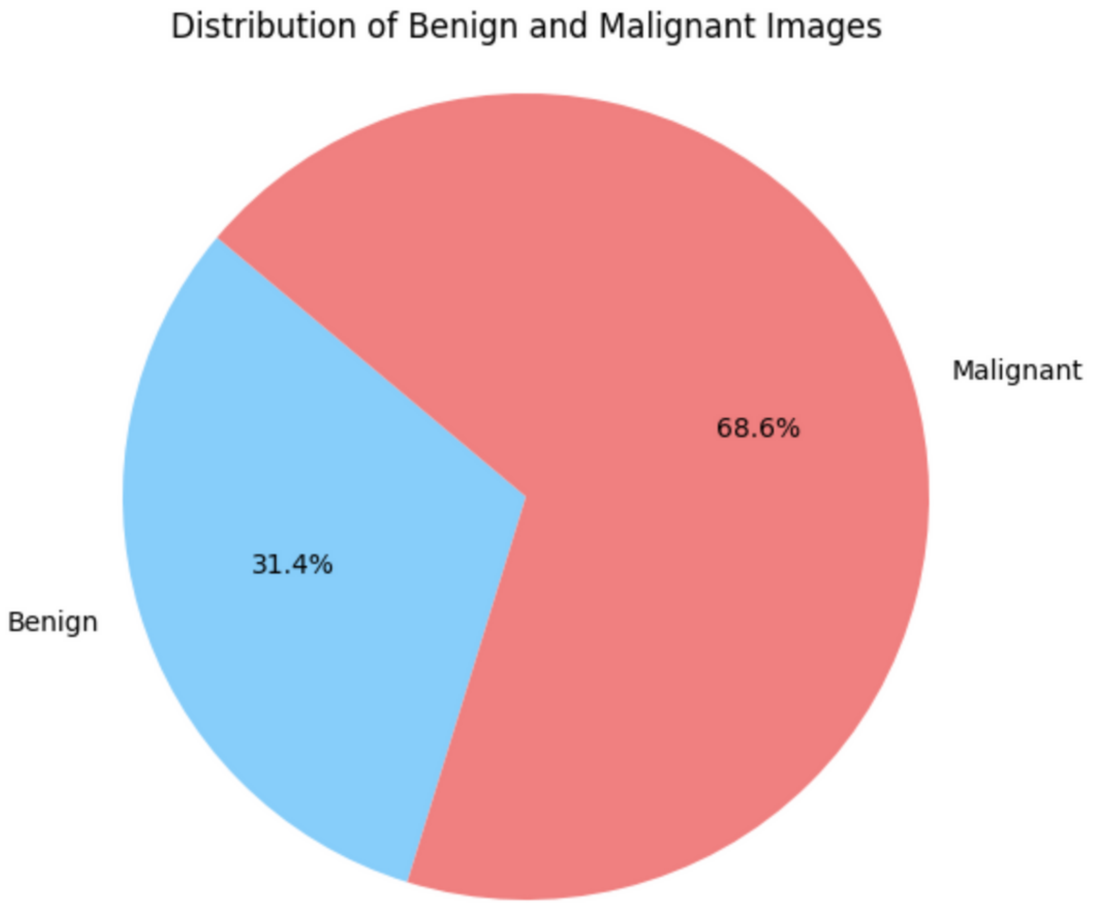
Exploratory Data Analysis

Sample Images

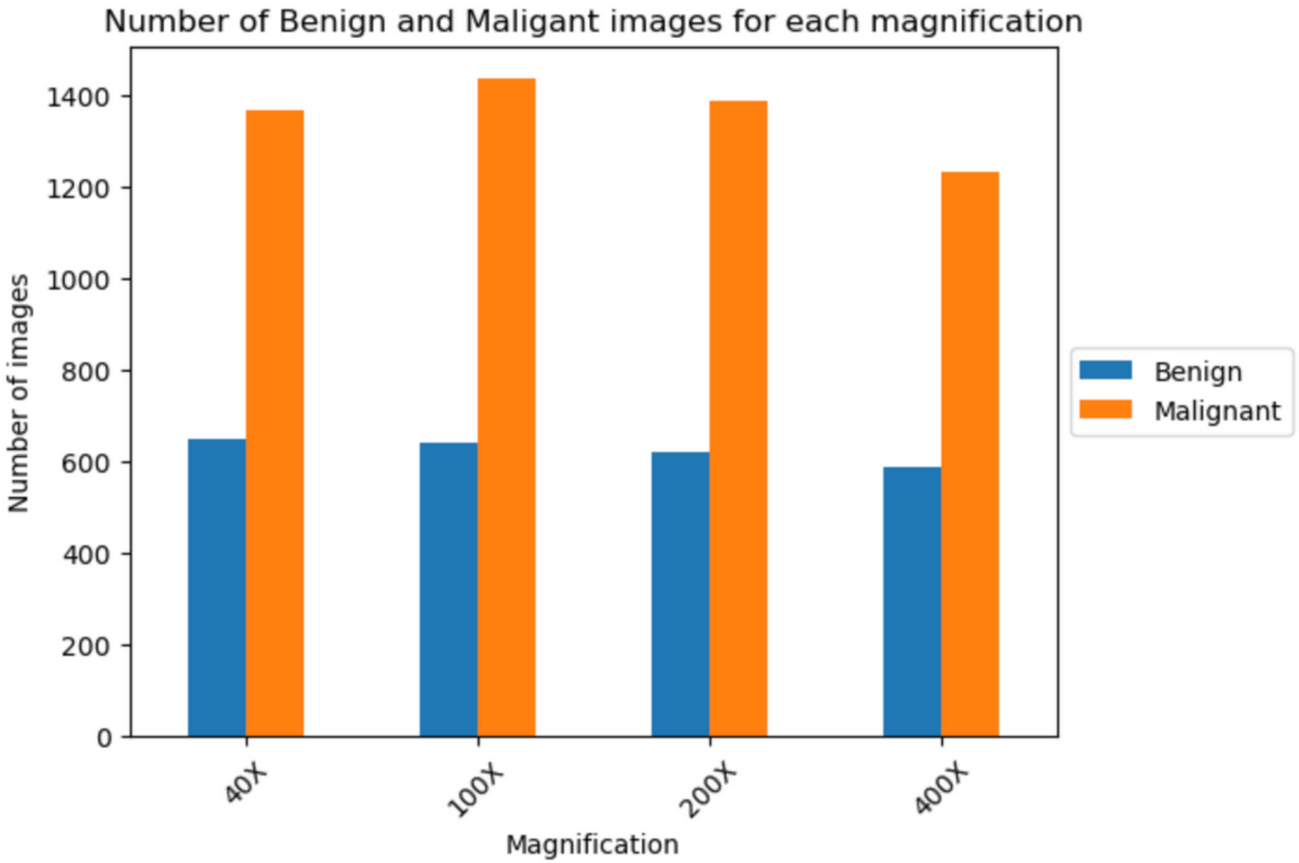




Class Distribution



Magnification Levels



Class Imbalance

- Accuracy is not a reliable metric
- Confusion Matrix
- ROC-AUC performance metric insensitive to class imbalance

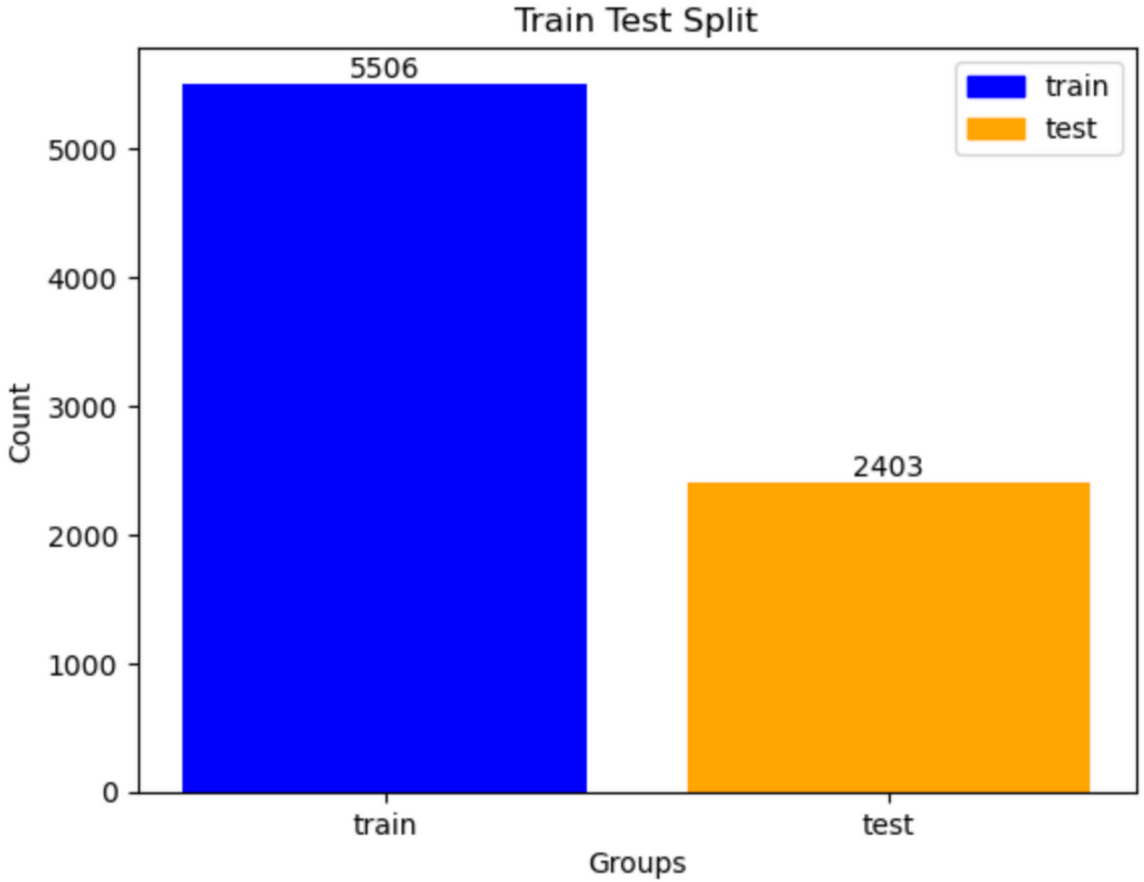
Magnification Levels

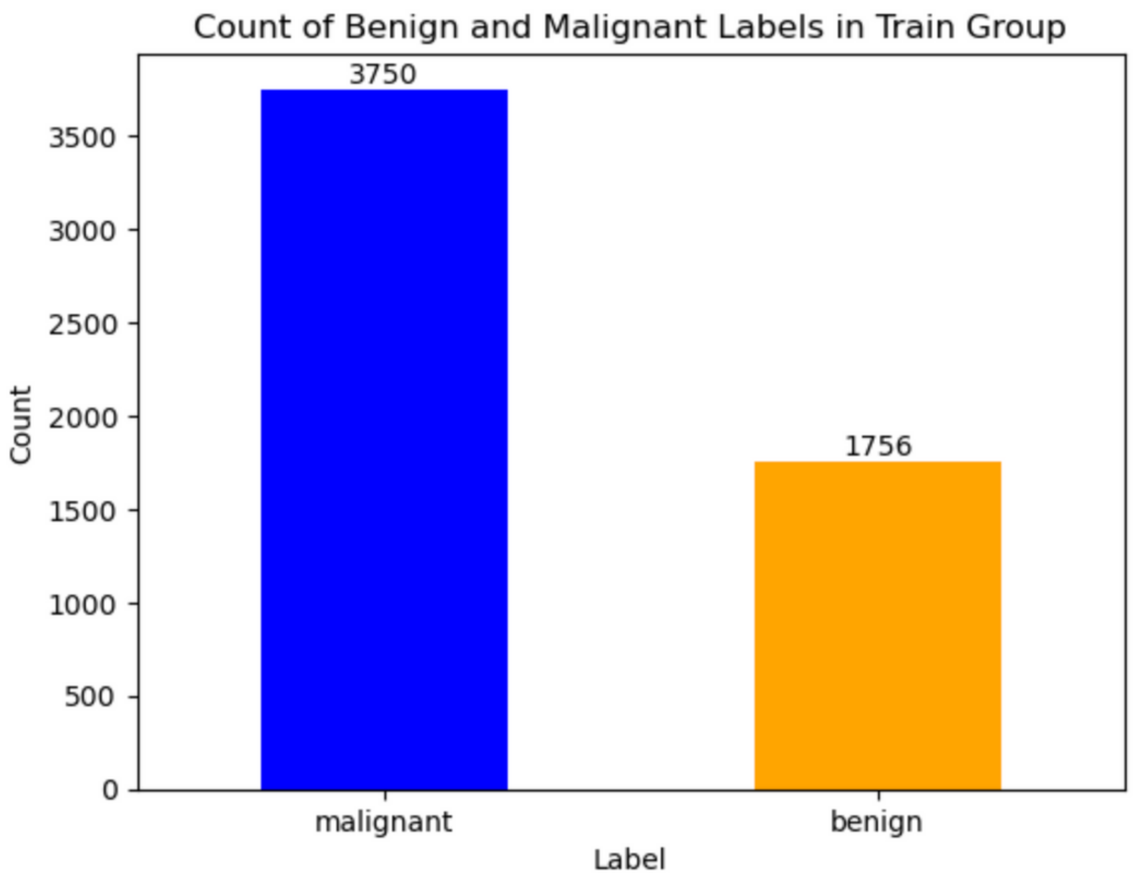
- 200x and 400x
- Improved model prediction

Data Preprocessing

Train Test Split

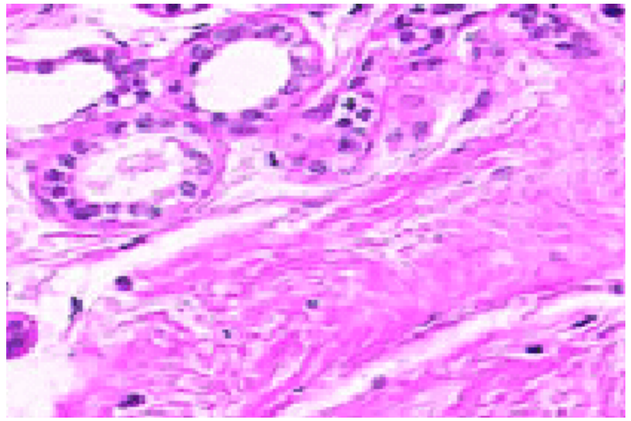
The BreaKHis dataset has been randomly divided into a training (70%) and a testing (30%) set.
To ensure the classifier generalizes to unseen patients, the patients used to build the training set are not used for the testing set.





Loading the images

Tensorflow library - To read and decode images into 3D tensors

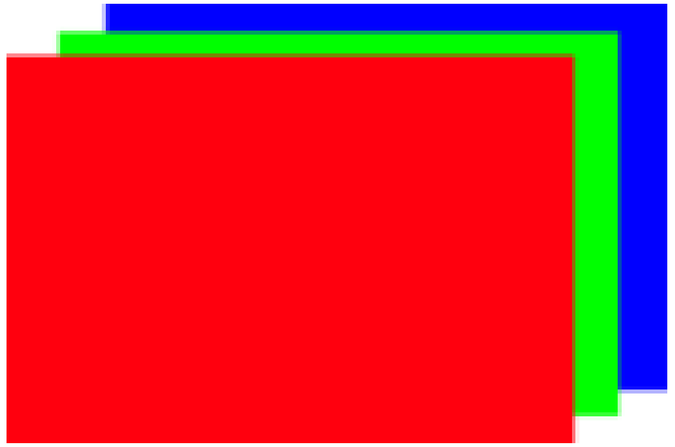


.PNG

tf.io.read_file(filename)

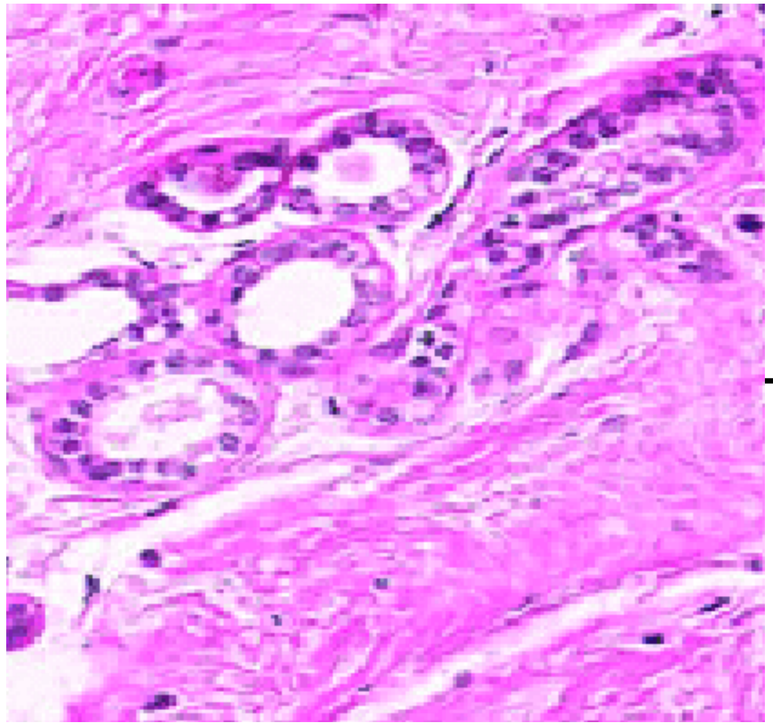
tf.image.decode_png(img_bytes)

IMAGE CONTENTS AS BYTES

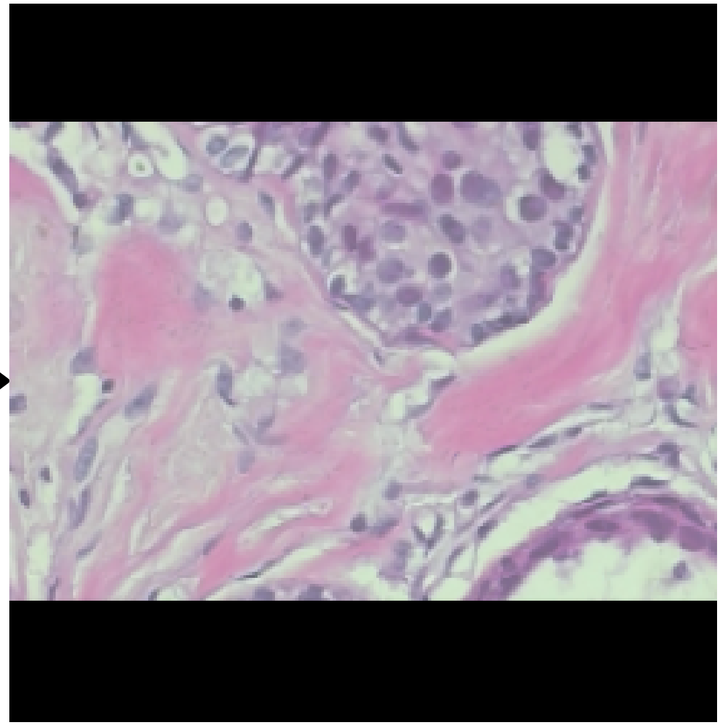


3D TENSOR

Resizing the images

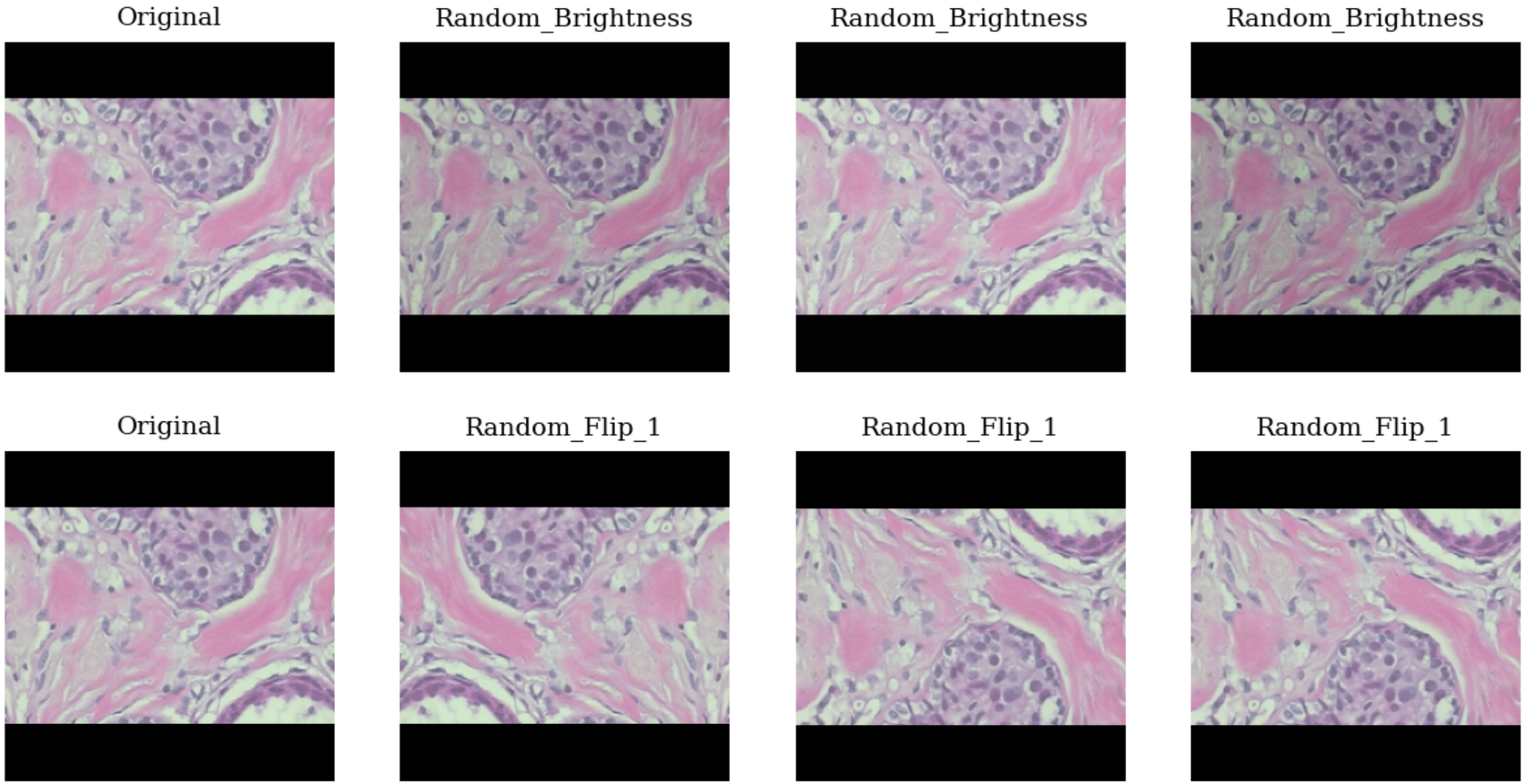


IMAGES OF VARIOUS DIMENSIONS

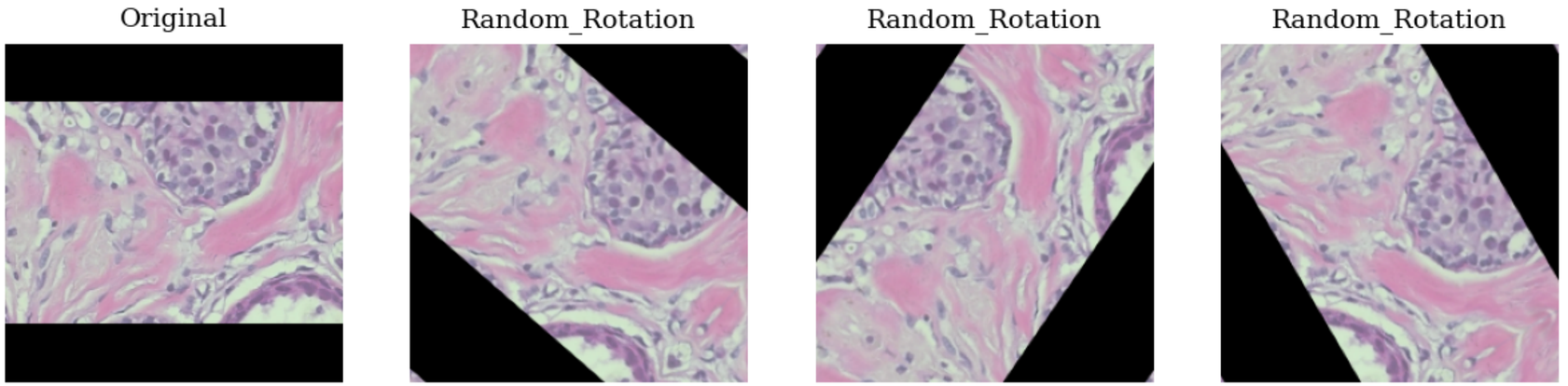


224 * 224 * 3

Data Augmentation

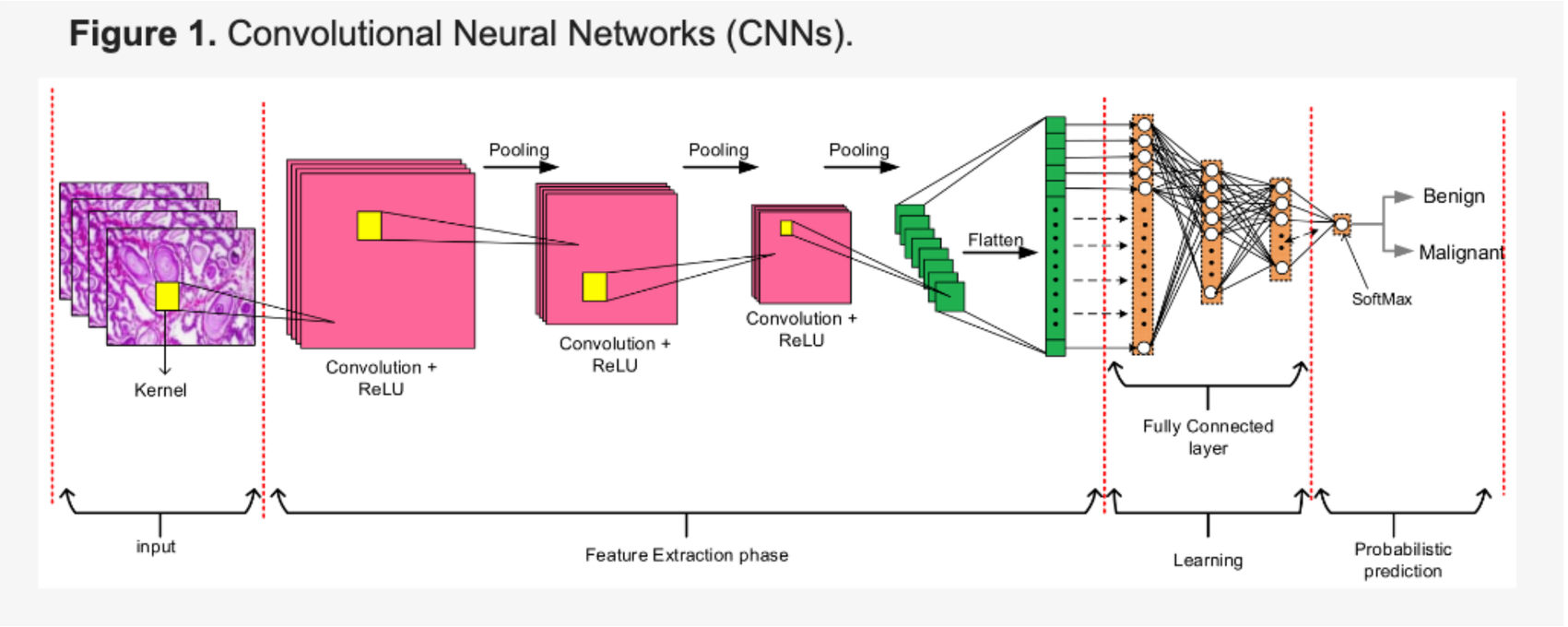


Data Augmentation



Computer Vision Models

Custom CNN using tensorflow.keras API



Model Metrics

ROC-AUC:	**0.85752**
Accuracy:	**0.83558**
Loss:	**0.50510**

ROC-AUC:

- **Area under the ROC curve. Measures a model's ability to distinguish between 2 classes**
- **Insensitive to class imbalance**
- **Ranges from 0.5 (random guesser) to 1 (perfect model)**

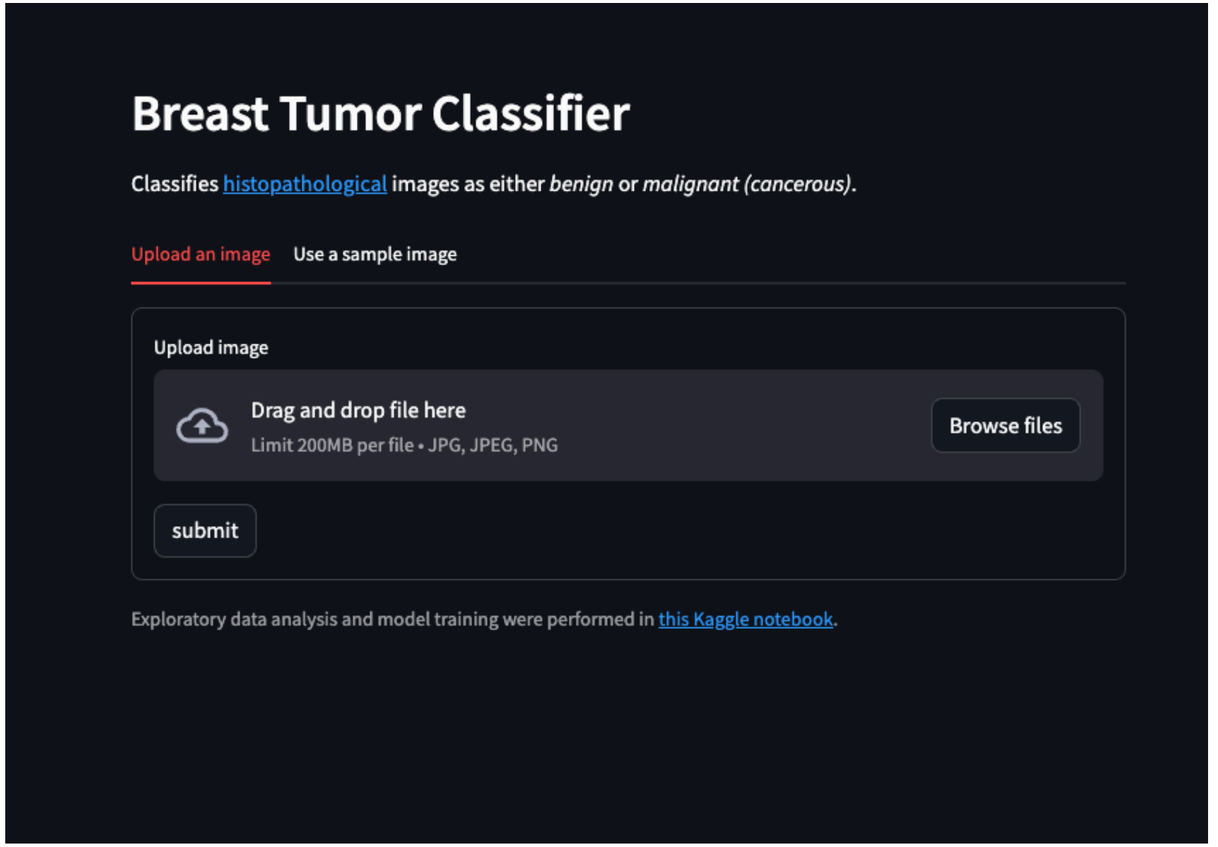
Accuracy:

- **Sensitive to class imbalance since a model can score highly just by predicting the majority class.**
- **Ranges from 0 (always wrong) to 1 (perfect model)**

Loss:

- **Places a heavy penaly on wrong predictions in order to direct the model to learn to predict correctly.**
- **Ranges from 0 (perfect model) to 1 (terrible model)**

Streamlit Demo App



Conclusion

The custom CNN model has demonstrated promising performance based on the following key metrics:

The model achieves an **ROC-AUC** score of 0.85752, indicating its ability to effectively distinguish between positive and negative instances.

The model achieves an **accuracy** of 0.83558, implying that approximately 83.56% of the model's predictions are correct. While accuracy provides a general overview of the model's performance, it may not fully capture its ability to balance true positives and true negatives, particularly in this imbalanced dataset.

The model's **loss value** stands at 0.50510, representing the average error during training. Lower loss values indicate better alignment between predicted and true labels. While the loss is relatively moderate, further optimization may be explored to minimize discrepancies between predictions and ground truth labels.