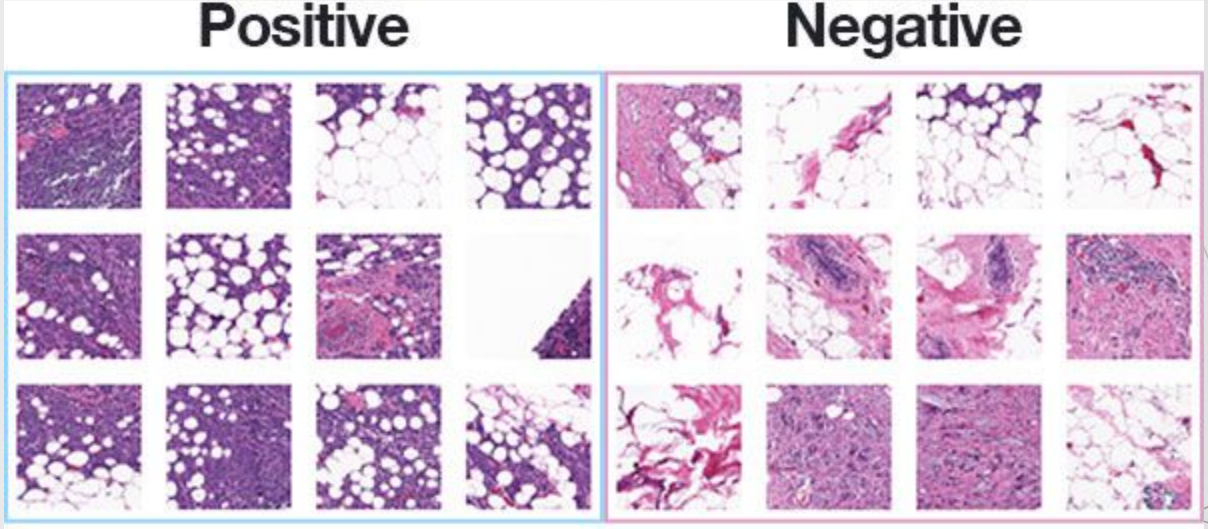
Histopathologic Cancer Detection

By: Dilsher Bhat and Timothy Hakobian

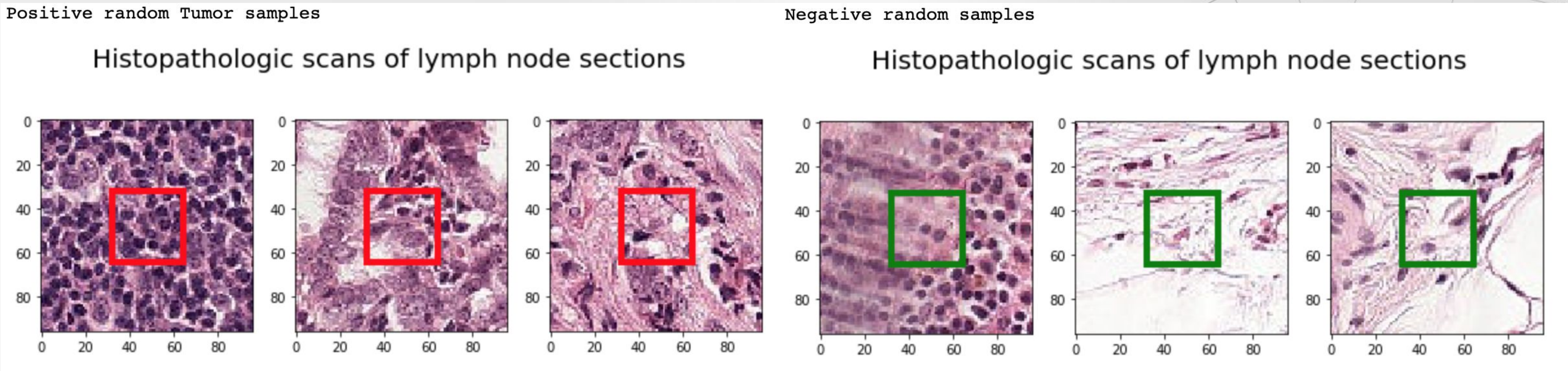
Background Information

- Histopathology: The diagnosis and study of diseases of the tissues, and involves examining tissues and/or cells under a microscope
- Histopathologic Cancer Detection involves finding cancer cells by looking at the tissues that may contain it under a microscope



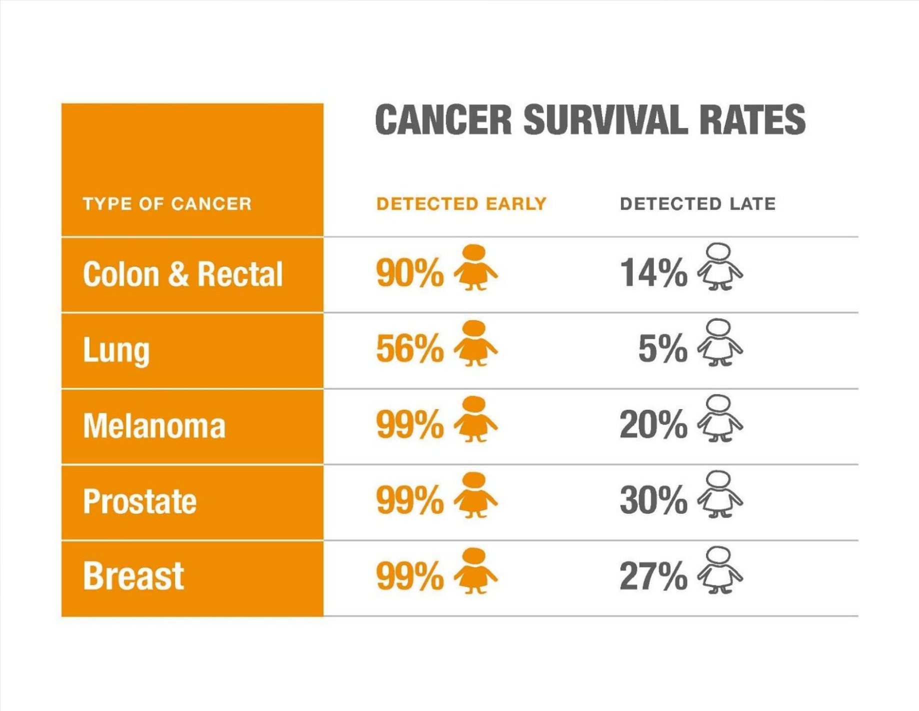
DataSet

- We use a modified version of the PCAM (PatchCamelyon dataset) which is available on kaggle: https://www.kaggle.com/c/histopathologic-cancer-detection/data
    - Number of images: 220025
- It is a unique dataset of annotated, whole-slide digital histopathology images of glass slide microscope images of lymph nodes that are stained with hematoxylin and eosin (H&E).
- This dataset is, in fact, a combination of two independent datasets collected in Radboud University Medical Center (Nijmegen, Netherlands), and the University Medical Center Utrecht (Utrecht, the Netherlands) who produced the slides by routine clinical practices and a trained pathologist would examine similar images for identifying metastases.


Positive random Tumor samples
Histopathologic scans of lymph node sections
Negative random samples
Histopathologic scans of lymph node sections
0
20
40
60
80

The Problem

- Detecting metastatic cancer is a challenging task for pathologists
- Pathologists have to look over large amounts of tissue to identify metastases, which have the possibility of being as small as single cells
- This process leaves room for a lot of errors for pathologists. Errors which can have detrimental effects on patients
- Early and accurate detection is crucial when it comes to treating cancer.



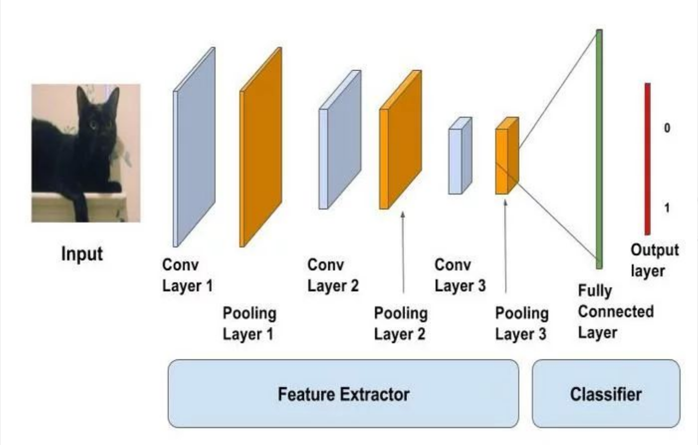
CANCER SURVIVAL RATES

TYPE OF CANCER	DETECTED EARLY	DETECTED LATE
Colon & Rectal	90%	14%
Lung	56%	5%
Melanoma	99%	20%
Prostate	99%	30%
Breast	99%	27%

From 15-40, a growing national organization devoted to improving cancer survival rates through early detection

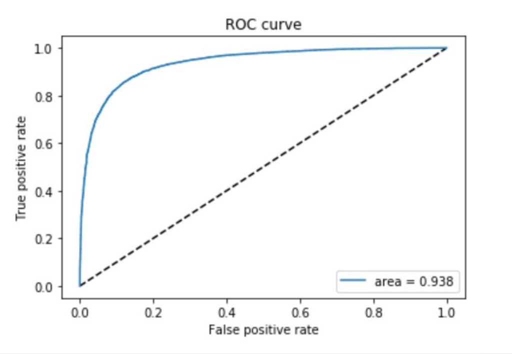
Our Approach

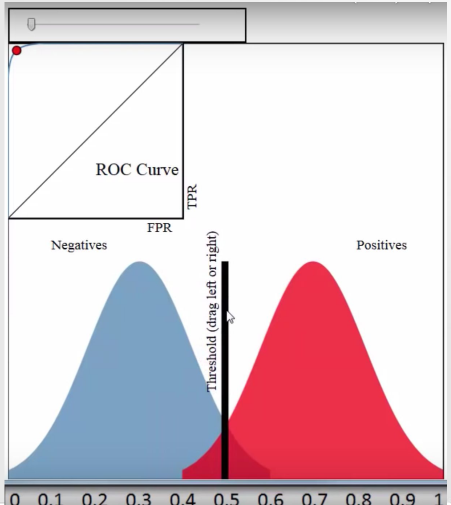
- We used openCV for pre-processing images and perform image augmentation in terms of geometric transformations such as axis flipping, color space changes, cropping and rotation
- After pre-processing our images, we used keras library to generate a convolutional neural network model based on binary image classification to identify metastatic cancer in the small image patches of the larger digital pathology scans.

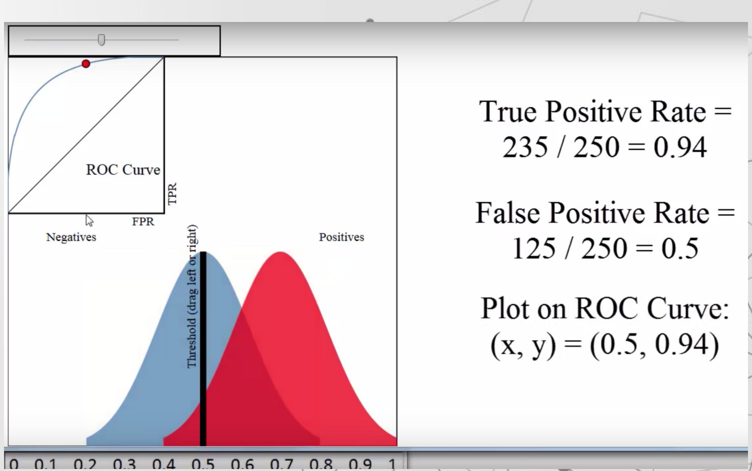


How We Evaluate Our Approach

- We evaluated our data by measuring the area under the receiver operating characteristic curve.
- ROC curves are often utilized to see the performance of a binary classifier. This means the classifier will have two possible outputs. A ROC curve is a good metric because it clearly shows the data pertaining to whether a patient has metastatic cancer or not.
- The y-axis of the curve represents the true positive rate and the x-axis represents the false positive rate. There is a threshold value that can be set at the place which divides the two possibilities.







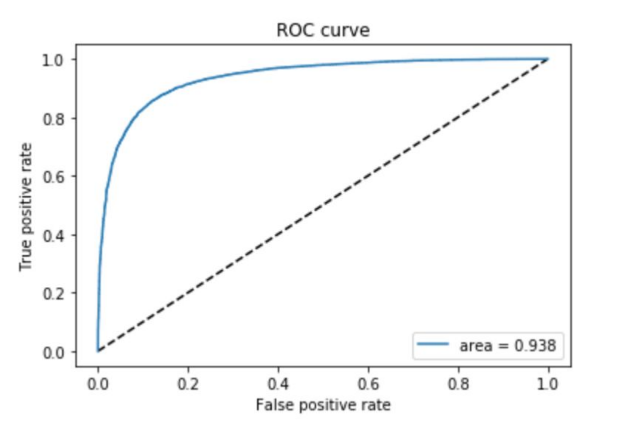
True Positive Rate = 235 / 250 = 0.94

False Positive Rate = 125 / 250 = 0.5

Plot on ROC Curve: (x, y) = (0.5, 0.94)

Discussion and Conclusion

- Our area under the curve is .938/1 which states that our model performs approx 93.8% of the classifications successfully to predict which scans have metastatic cancer.
- Even though the percentage of successful predictions is a high 93%, in terms of application to real life detection, such a model performs poorly since the pathologist's decision of whether a scan has cancer or not can result in saving a life, or incorrectly judging their disease and threatening the person's life further.
- Thus, we will strive to increase the percentage of successful predictions to the highest it can be (around 99.9%) by improving pre-processing of the dataset and machine learning model that the data is being trained on. Currently we're using a Sequential model to get results, however, we plan to delve into other models as well outside the keras library.


ROC curve
True positive rate
False positive rate
area = 0.938