

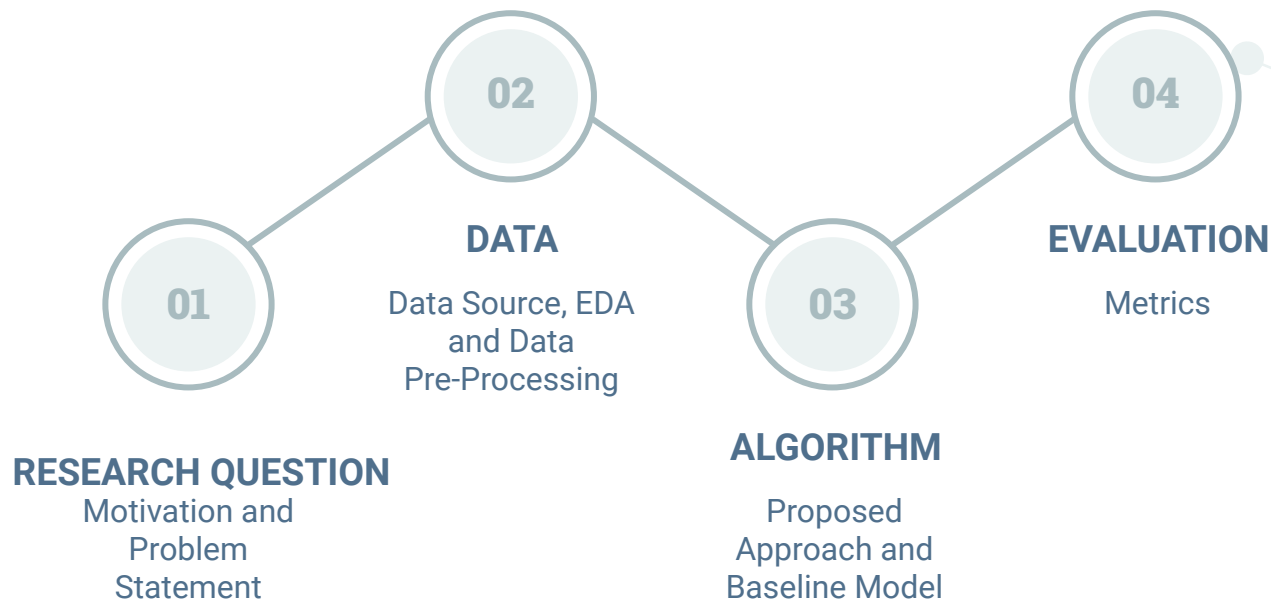


# Tumor Detection From Histopathological Slides

Names: Srila Maiti, Justin To, Hector Rincon, Chenyu Wang and Ifrah Javed

---

# Overview



A decorative network diagram consisting of numerous dark blue circles of varying sizes connected by thin, light blue lines, forming a complex web-like structure. A vertical white line extends from the top of the slide down to a central circular node. This node is a white circle with a thin grey border, containing the number '01' in a dark blue font. The entire graphic is set against a solid dark blue background.

01

# Research Question

Motivation and Problem Statement

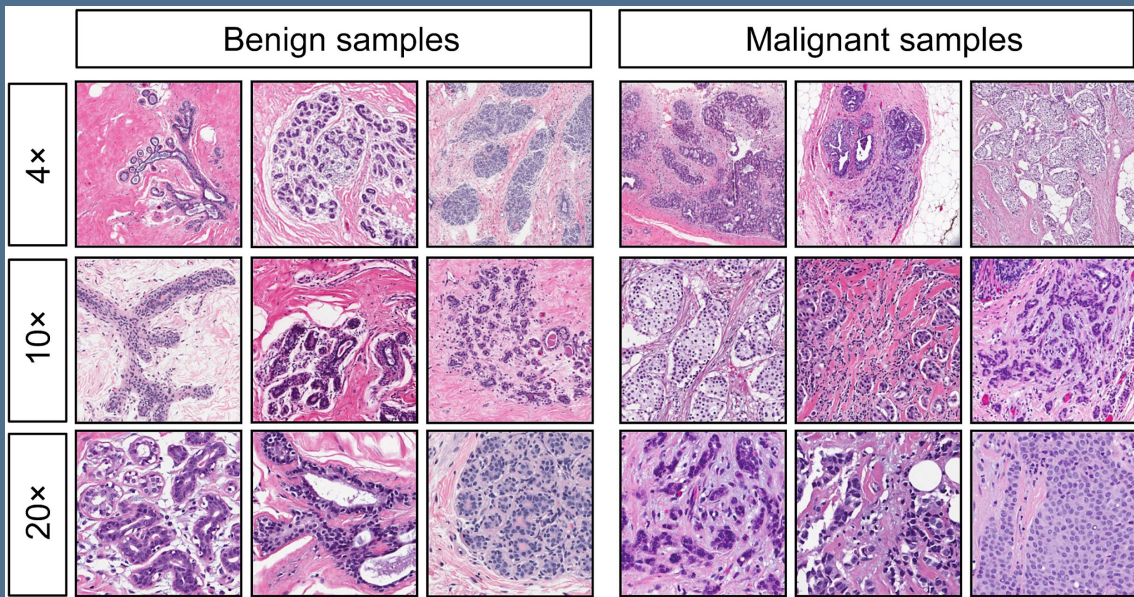
Cancer is the Rapid  
Growth of Mutated Cells

Histopathology is a  
Diagnostic Tool

28% Misdiagnosis Rate<sup>[1]</sup>

Early Diagnosis is Key

# Background





**Can we identify the presence of  
metastatic tumor from  
histopathological slides?**





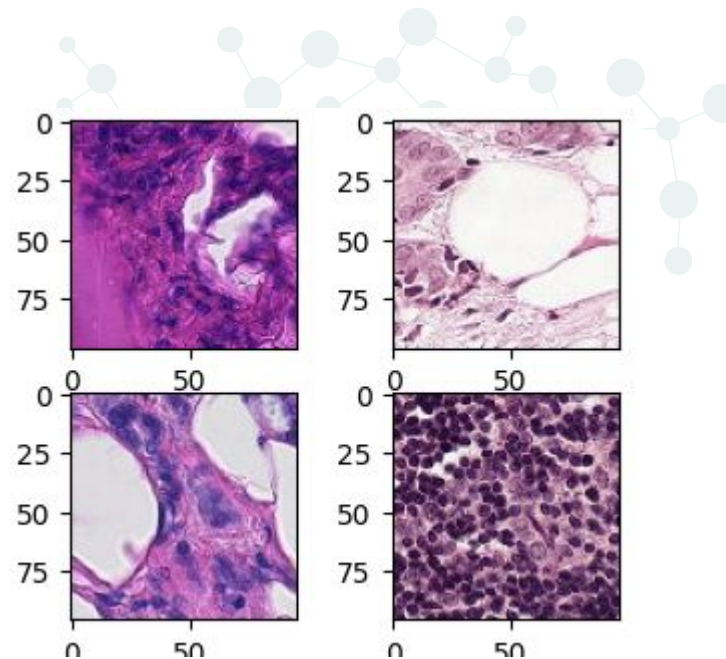
02

# Dataset

Data Source, EDA and Transformations

# Data

- Working with image data of pathological cells for cancer detection
- A version of the PCAM data set (<https://github.com/basveeling/pcam>)
- Cleaned from duplicates by Kaggle for the [Histopathologic Cancer Detection competition](#)
- Dataset contains 277,483 total images (220,025 train / 57,458 test)
- Positive label: center 32x32px (out of 96x96 image) region of a patch contains at least one pixel of tumor tissue

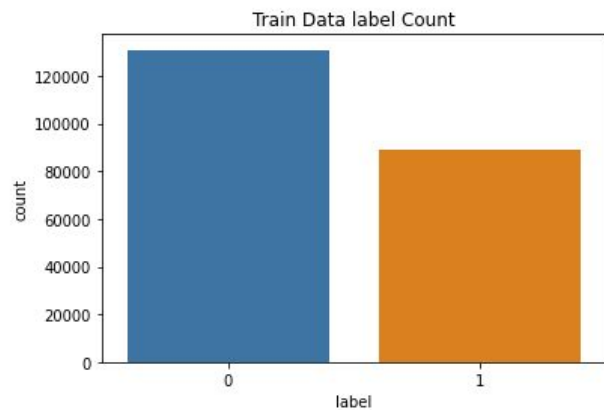


[1] B. S. Veeling, J. Linmans, J. Winkens, T. Cohen, M. Welling. "Rotation Equivariant CNNs for Digital Pathology". [arXiv:1806.03962](https://arxiv.org/abs/1806.03962)

[2] Ehteshami Bejnordi et al. Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer. JAMA: The Journal of the American Medical Association, 318(22), 2199–2210. [doi:jama.2017.14585](https://doi.org/10.1001/jama.2017.14585)

# Data

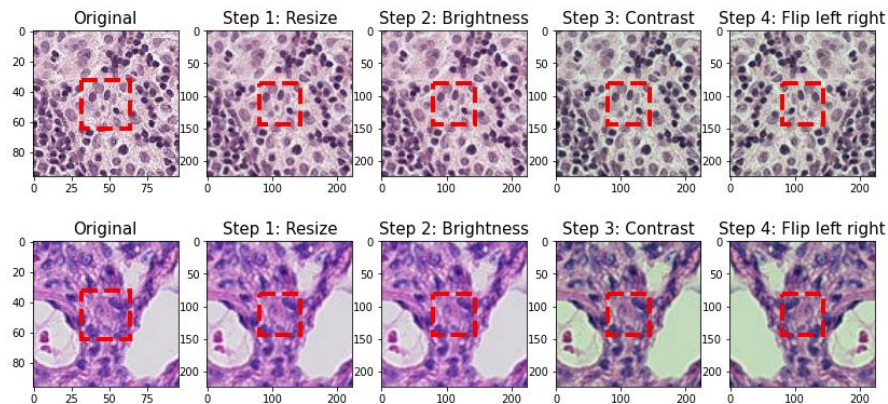
There is a slight class imbalance between the positive and negative labels in the train data. To remediate this, we will undersample from the negative label class





# Data Preprocessing

- Images were:
  - Converted to grayscale
  - Resized randomly
  - Standardized brightness
  - Standardized contrast
  - Rotated 90, 180, and 270 degrees
  - Randomly flipped
- In the train data there are:
  - 130,908 negative (59.5%)
  - 89,117 positive examples (40.5%)
- We will use a 60/20/20 split (train, validation, test)



A decorative network diagram consisting of numerous dark blue circles of varying sizes connected by thin, light blue lines, forming a complex web-like structure across the slide. A vertical white line descends from the top center, ending at a white circle with a dark blue border.

03

# Algorithm

Proposed Approaches

# Algorithm Approaches:



Logistic  
regression

CNN

# Algorithm: Logistic Regression

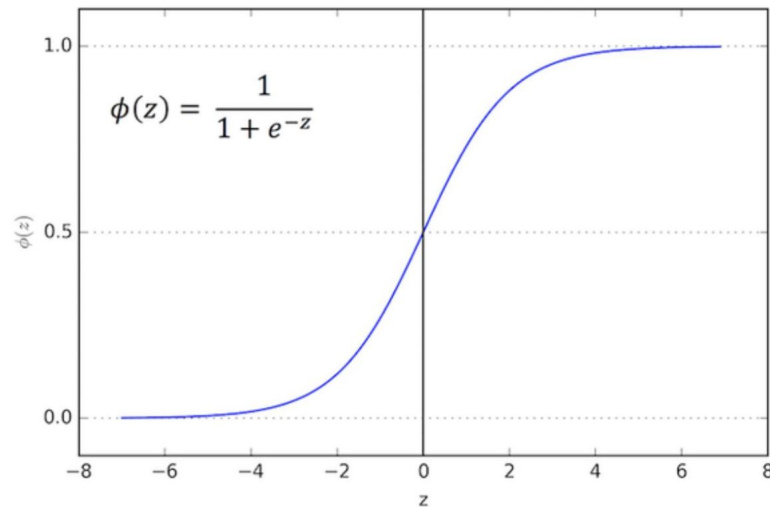
Problem category: Image Classification.

Output y: 0 or 1 (indicating whether there is at least one pixel of tumor tissue)

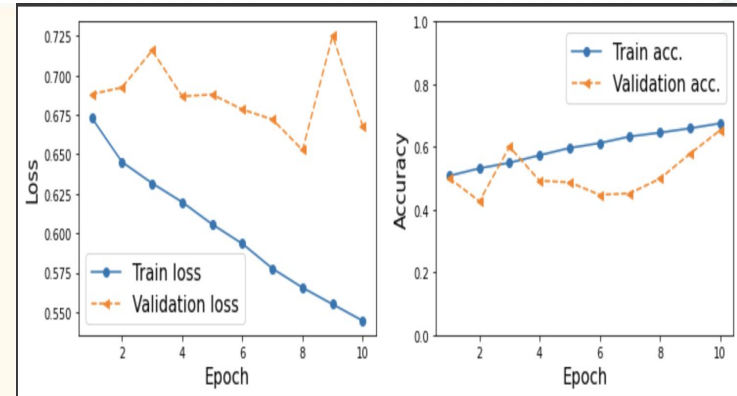
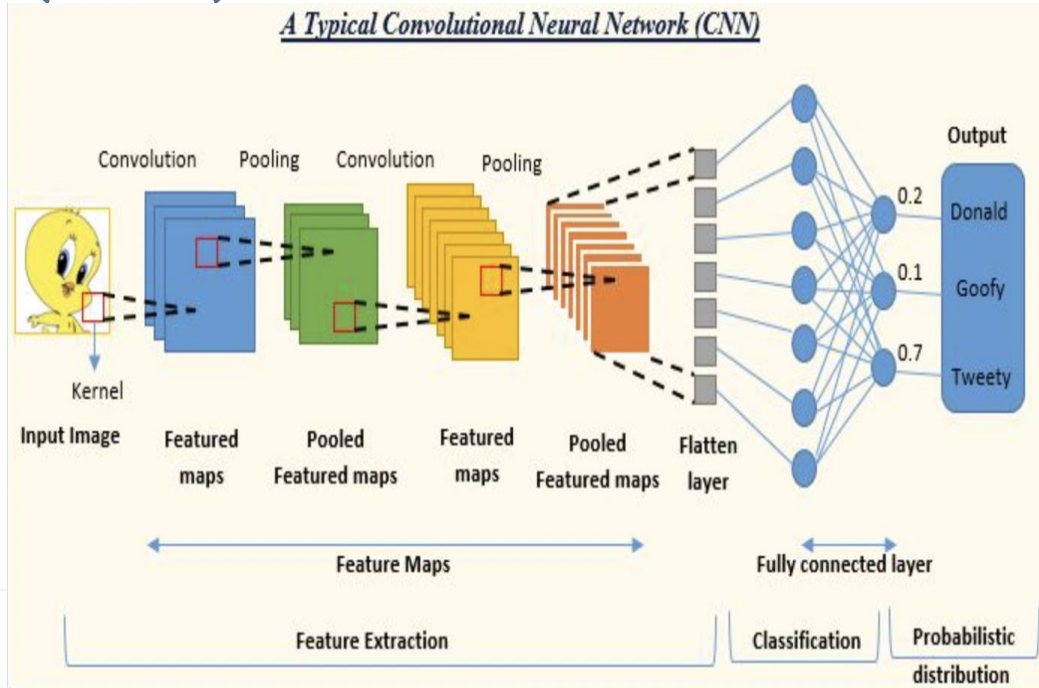
Input x:  $32*32 \rightarrow 1024$  total pixel.

Loss function: for logistic regression:

$$-\frac{1}{|Y|} \sum_{y_i \in Y} y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)$$



# Algorithm: Convolutional Neural Network (CNN)



Test Accuracy is 81.8% in the initial run.



04

# Evaluation

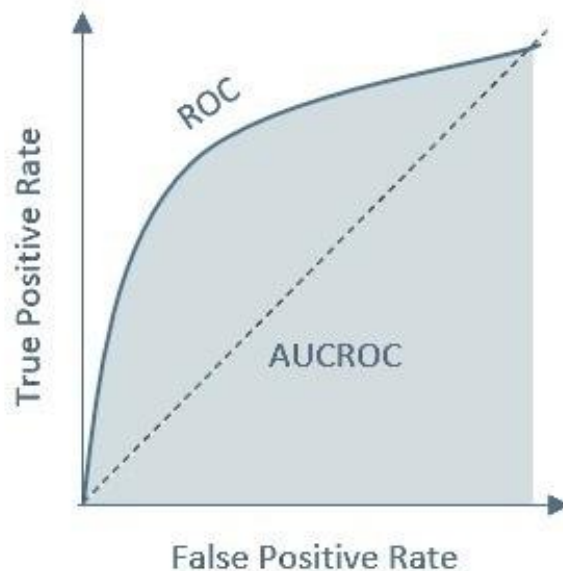
Metrics

# Model Evaluation

As the task is one of **classification**...

## Overall metric

- **AUC under ROC**
- general sense of model effectiveness
- yet to cater for use case
- **Baseline: 50%** (given lack of context)



# Model Evaluation

As the task is one of **classification**...

## Overall metric

- **AUC under ROC**
- general sense of model effectiveness
- yet to cater for use case
- **Baseline: 50%** (given lack of context)

## Additional metrics

- **Confusion matrix**
- **Precision, recall**: for specific use cases
- **F1, MCC**: balanced estimate



Predicted

True Label			
		Positive	Negative
Predicted	Positive	True Positives	False Positives
	Negative	False Negatives	True Negatives



# Model Evaluation

As the task is one of **classification**...

## Overall metric

- **AUC under ROC**
- general sense of model effectiveness
- yet to cater for use case
- **Baseline: 50%** (given lack of context)

## Additional metrics

- **Confusion matrix**
- **Precision, recall**: for specific use cases
- **F1, MCC**: balanced estimates

## Validation

- **k-fold** cross validation
- stretch goal: subject to computation time



		True Label	
		Positive	Negative
Predicted	Positive	True Positives	False Positives
	Negative	False Negatives	True Negatives



# Thank You

Questions?

---

