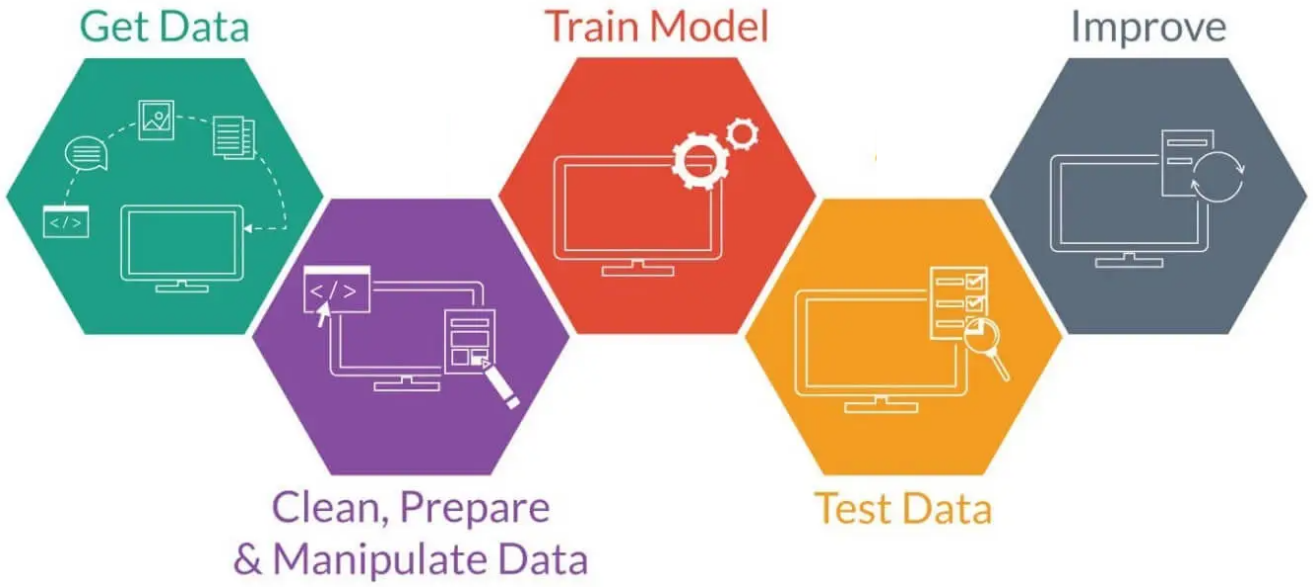
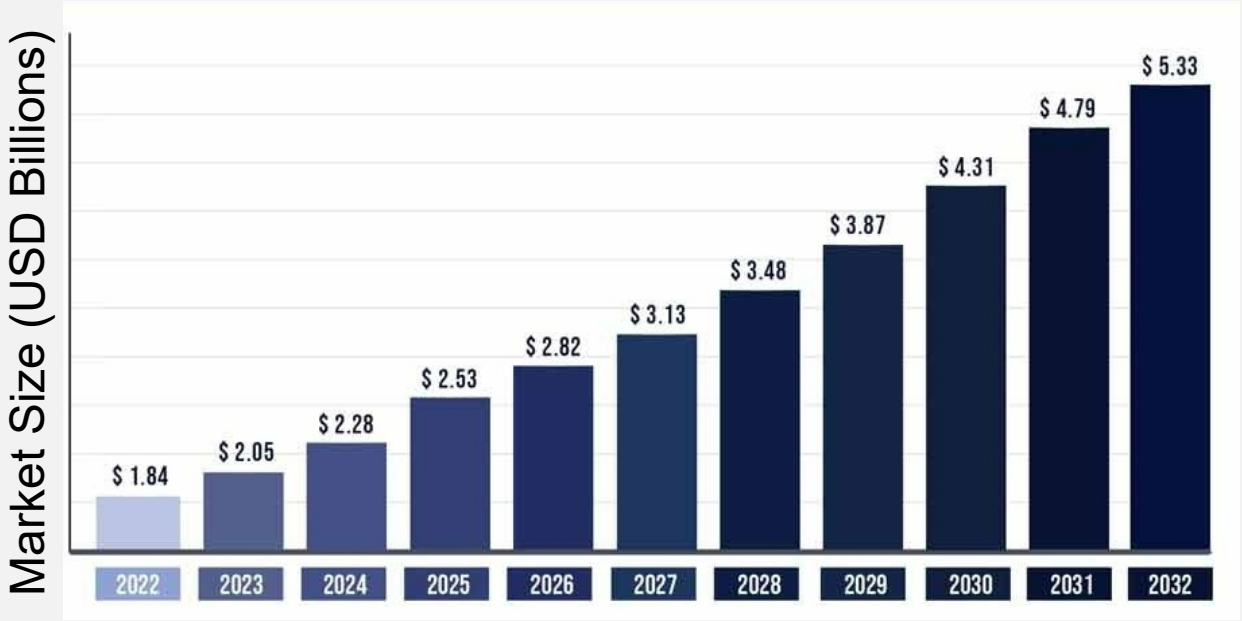
ML IN DIGITAL PATHOLOGY

Yuan Nghiem
Leo Grant Berman
Albert Bulik



PROBLEM STATEMENT

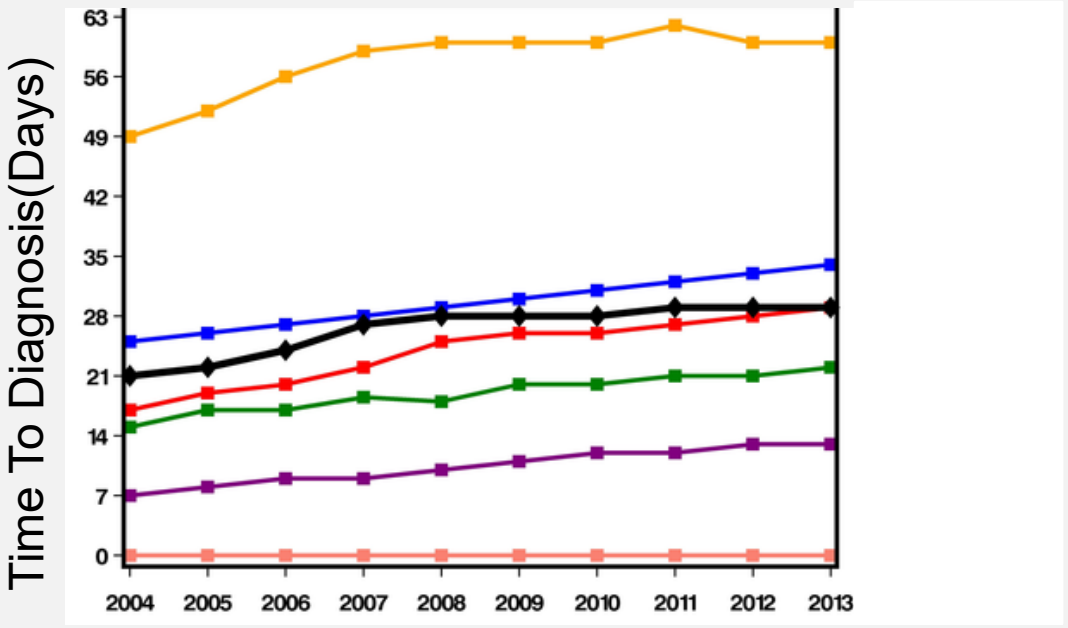
The demand for cancer diagnosis is increasing.



YEAR (2022 to 2032)

SOURCE

As a result, time to treatment is increasing.

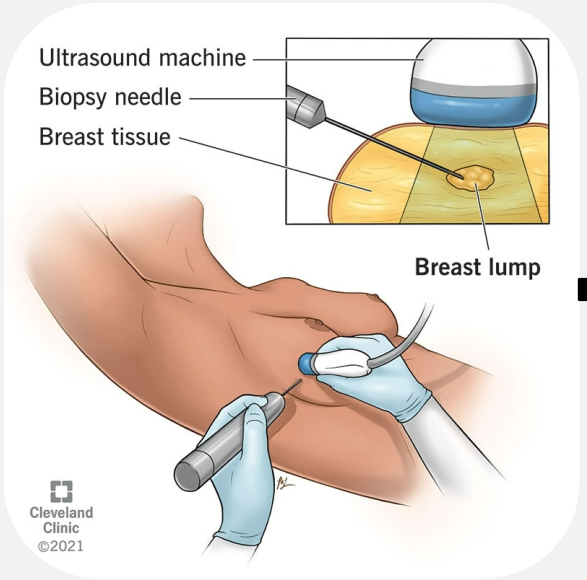


DIAGNOSIS YEAR (2004 to 2013)

SOURCE

TEMPLE UNIVERSITY HEALTH DIGITAL PATHOLOGY (TUHDP) CORPUS

Biopsy

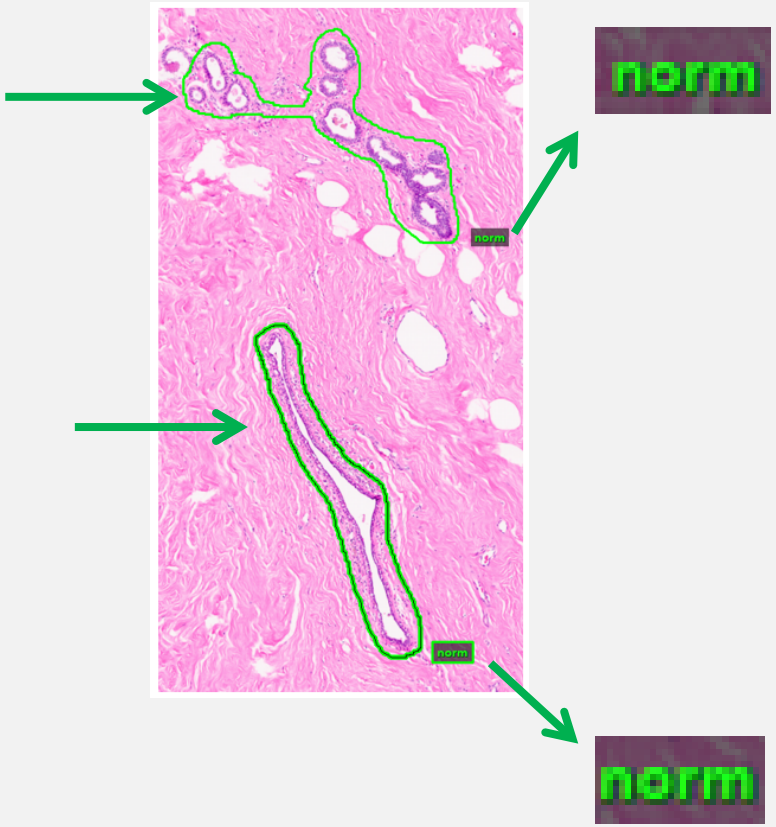


Staining

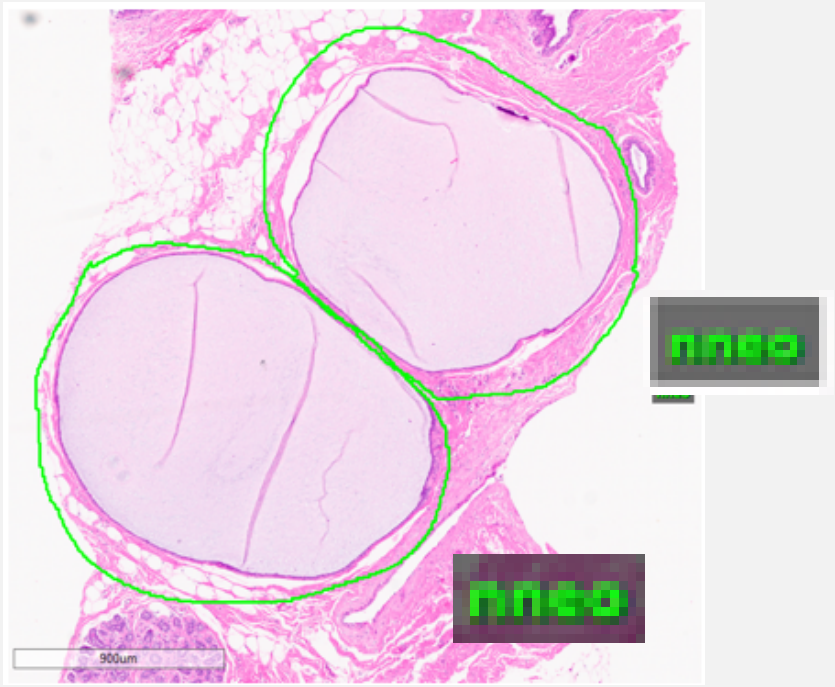
Scanning

LABELED DATA TYPES

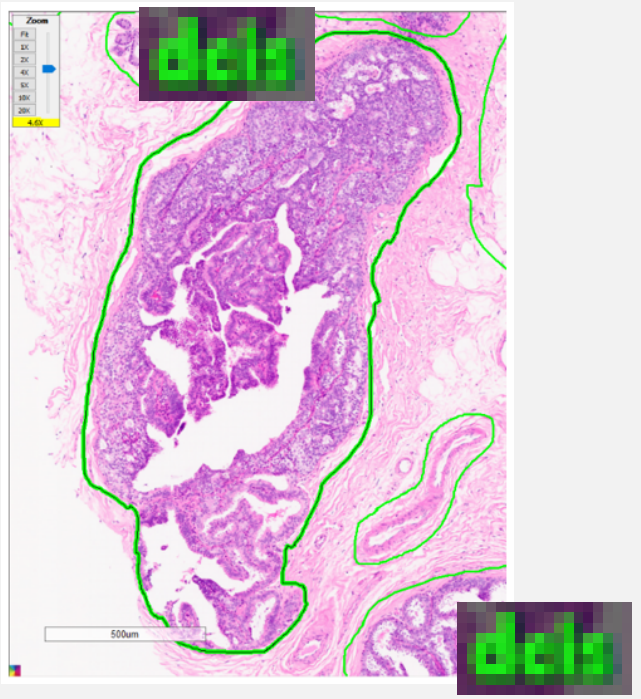
Non-Cancerous Types



Carcinogenic-Signs Type



Cancerous Types

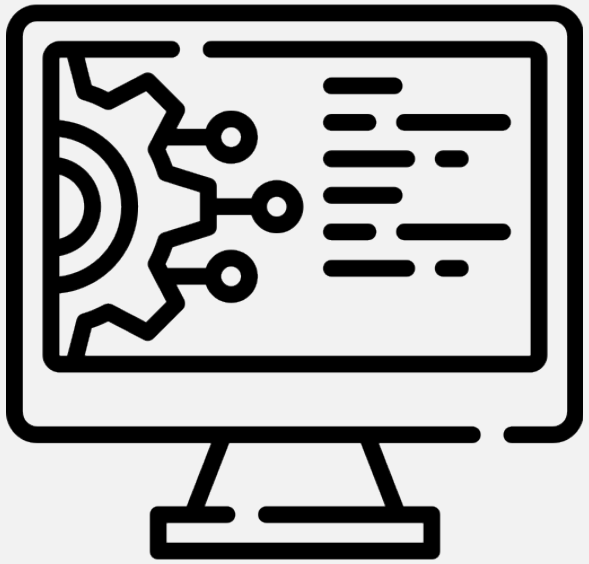


OUR PRODUCT

Scanning

(Data)

System



(Model)

Prediction Report

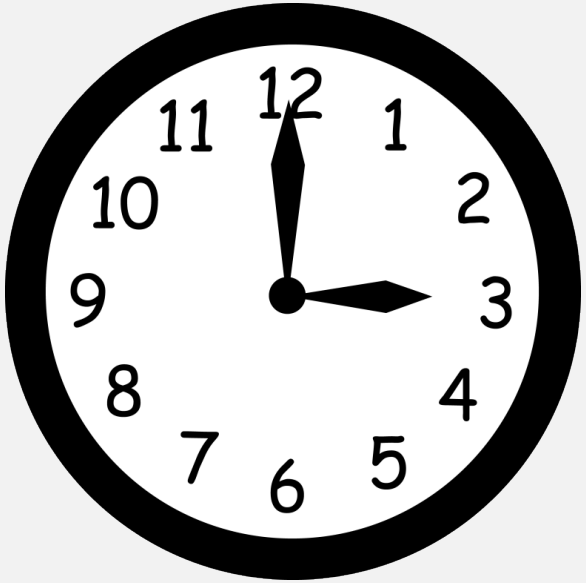
(Output)

DESIGN CRITERIA / REQUIREMENTS

Whole Slide Image Classification
Source

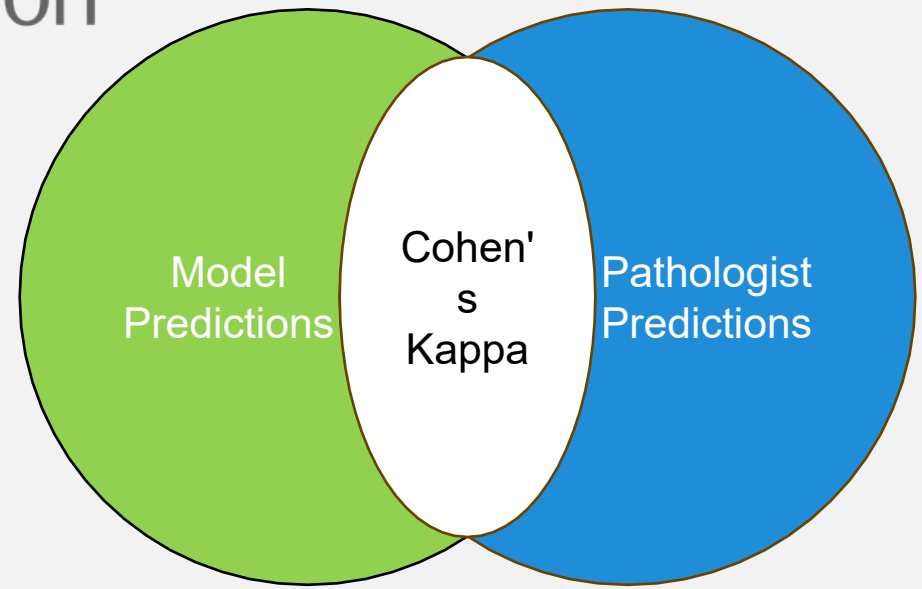
Functional GUI

- Show location of areas
- Show area's probability of malignancy



Train Time < 3 Days

DESIGN CRITERIA / REQUIREMENTS



Cohen's Kappa > 60%
Source

F1 Score

$$\frac{TP}{TP + .5(FN)(FP)}$$

F1 Score > 90%
Source

DEEPER LOOK AT F1 SCORE

Actual Classes

Predicted Classes

F1 Score

$$\frac{TP}{TP + .5(FN)(FP)}$$

Apples

True Positive = 4

False Negative = 0

False Positive = 1

F1 Score = 89%

Bananas

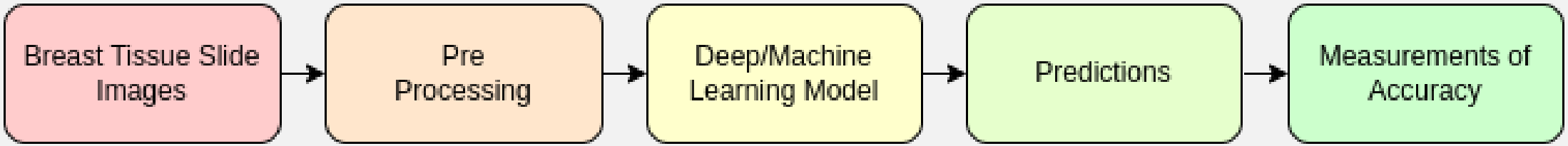
True Positive = 0

False Negative = 1

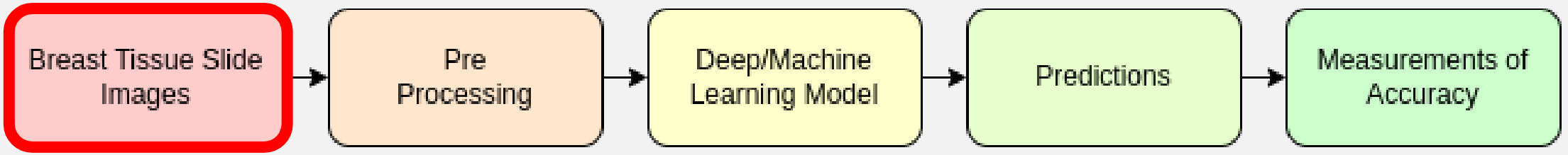
False Positive = 0

F1 Score = 0%

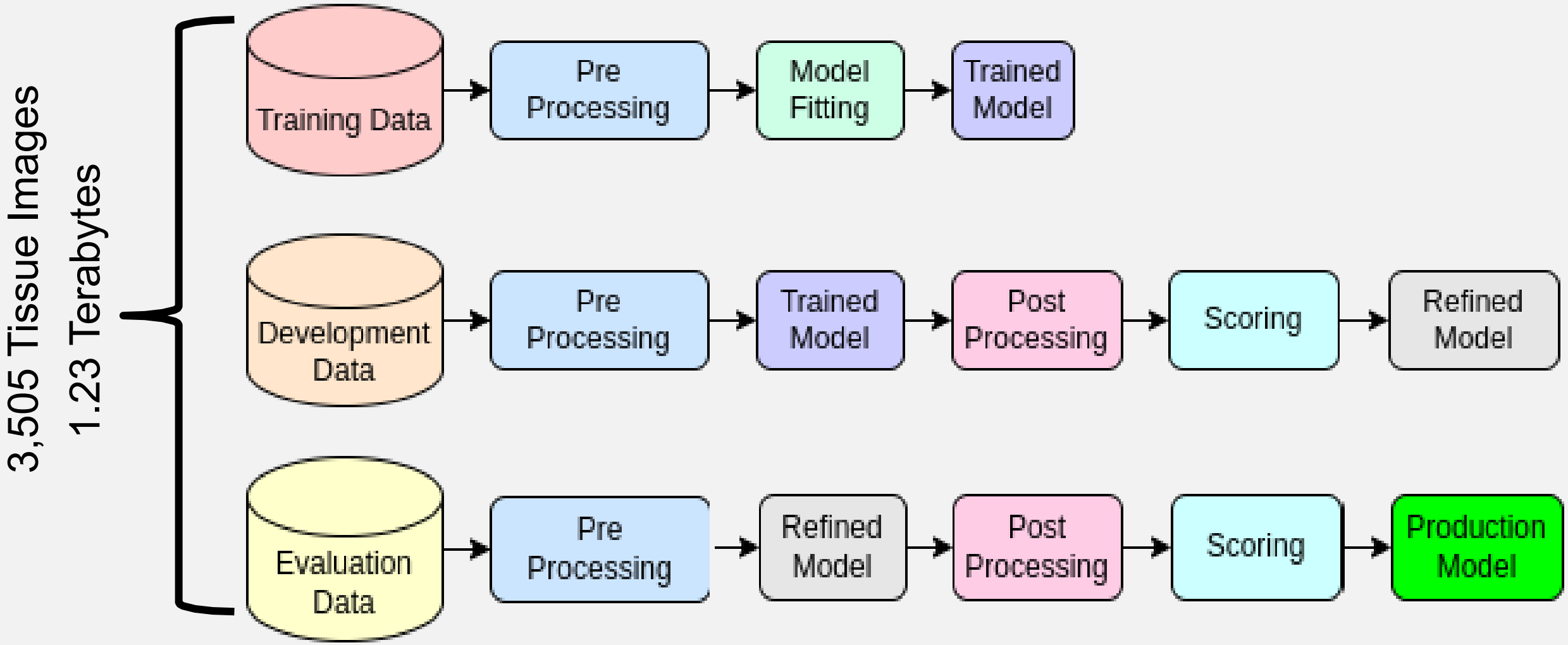
PRELIMINARY DESIGN



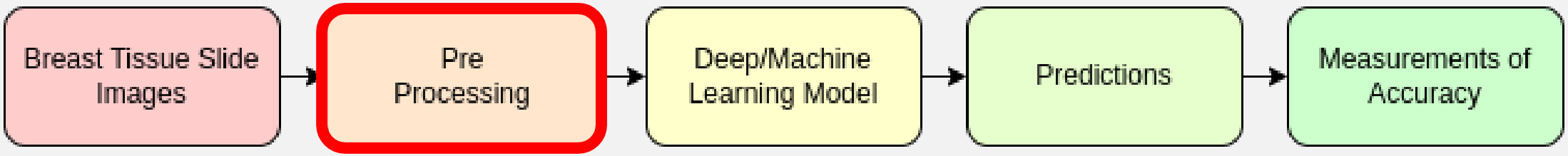
PRELIMINARY DESIGN



TRAINING, TUNING, EVALUATION PIPELINES



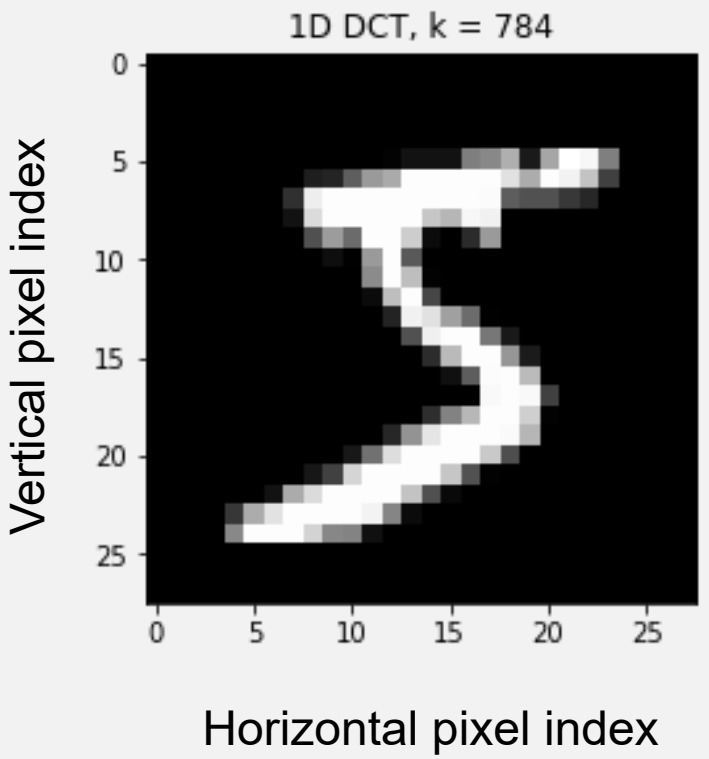
PRELIMINARY DESIGN



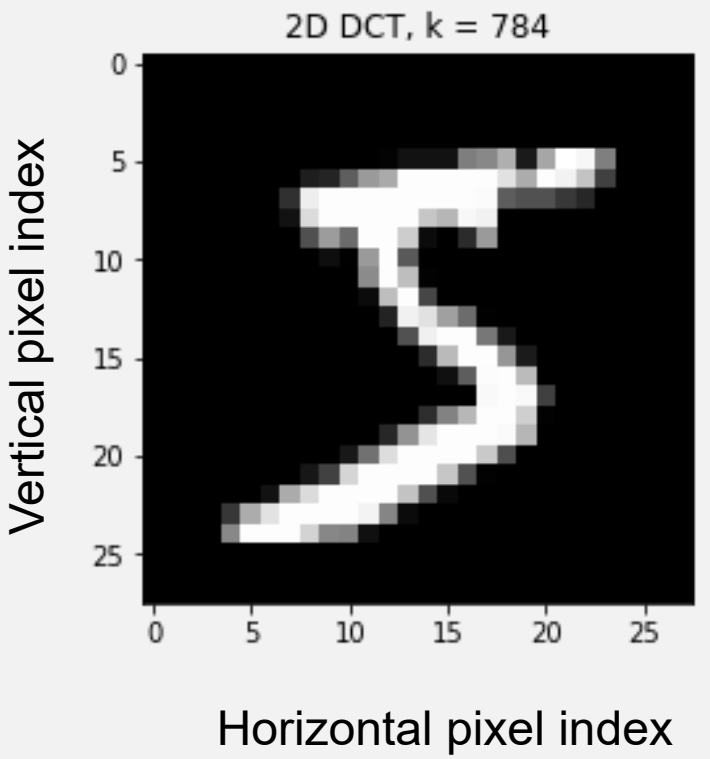
2-DIMENSIONAL DCT

- Belongs in the pre-processor stage
- Converts colors to frequencies
- Allows us to retain fewer features

One-dimensional DCT



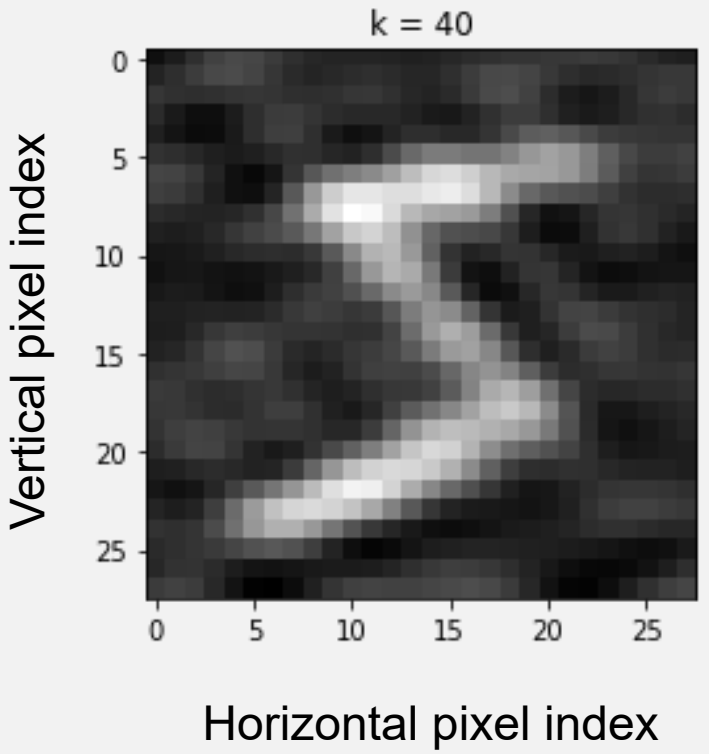
Two-dimensional DCT



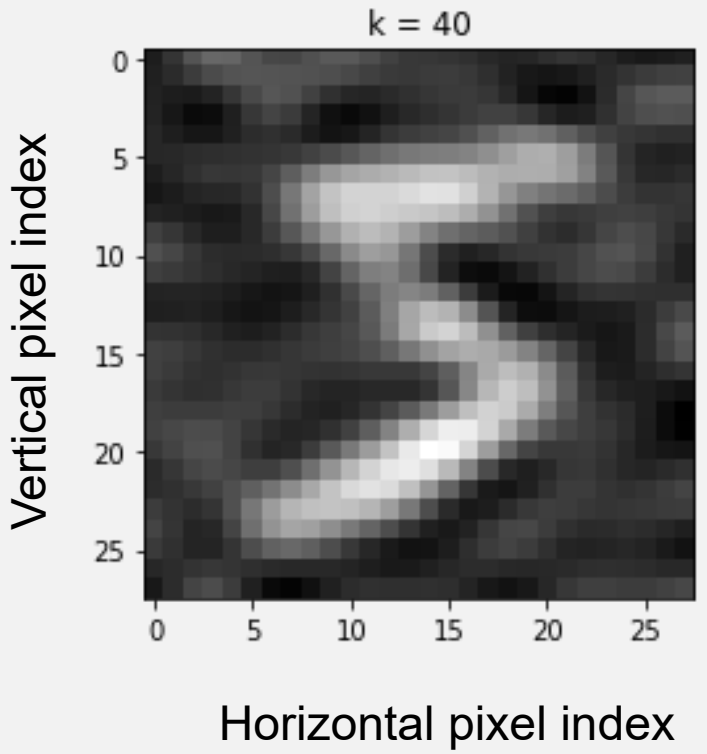
2-DIMENSIONAL DCT

- Belongs in the pre-processor stage
- Converts colors to frequencies
- Allows us to retain fewer features

One-dimensional DCT



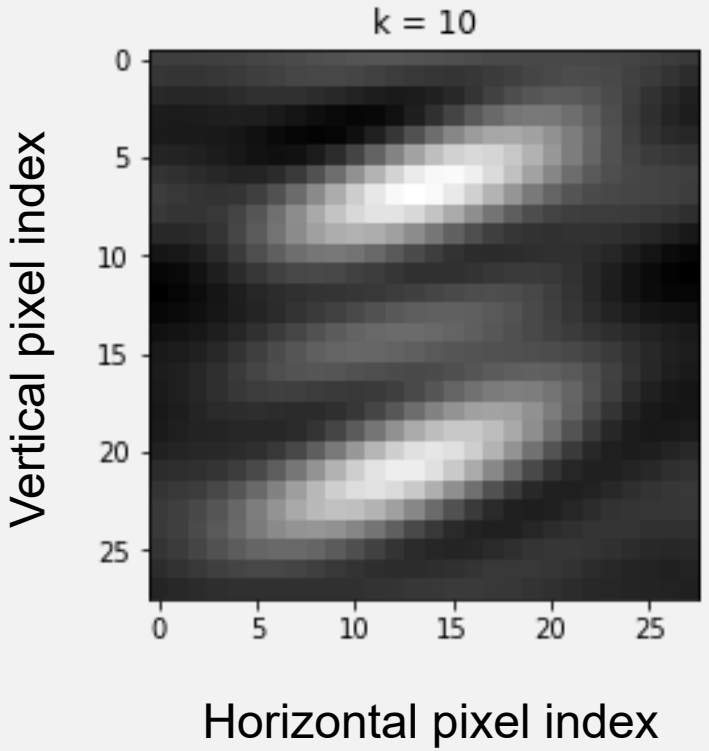
Two-dimensional DCT



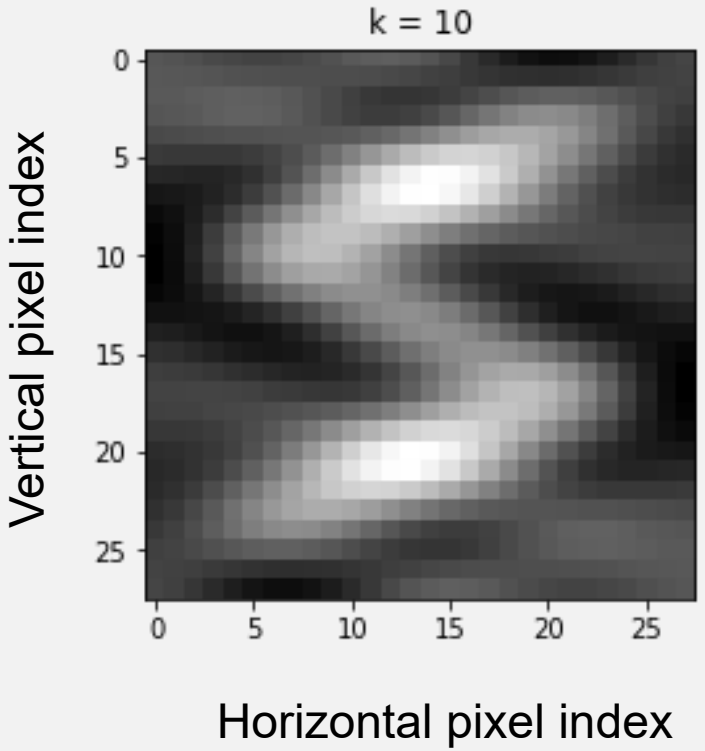
2-DIMENSIONAL DCT

- Belongs in the pre-processor stage
- Converts colors to frequencies
- Allows us to retain fewer features

One-dimensional DCT

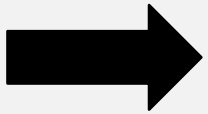


Two-dimensional DCT



PRINCIPAL COMPONENT ANALYSIS (PCA)

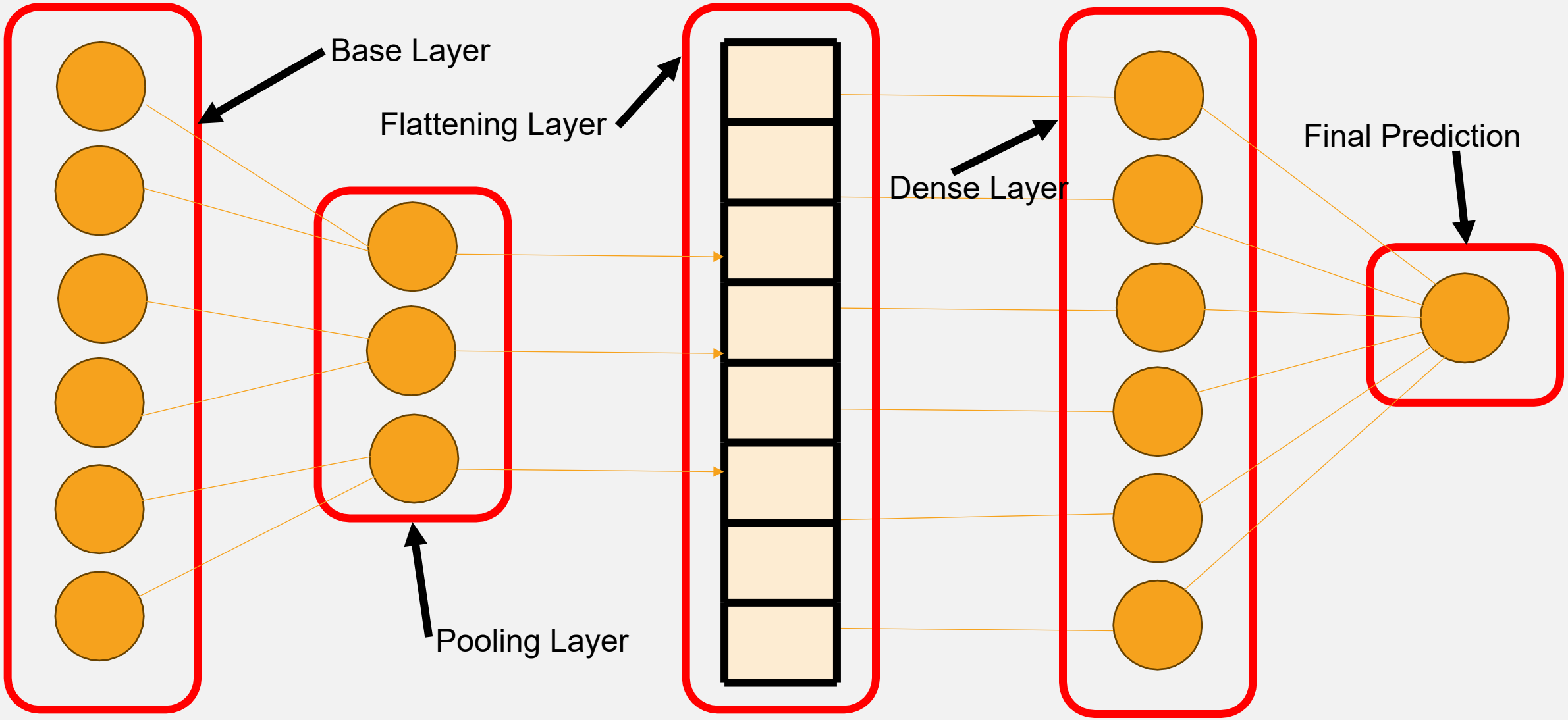
f1	f2	f4	f7
f3	f5	f8	f11
f6	f9	f12	f14
f10	f13	f15	f16



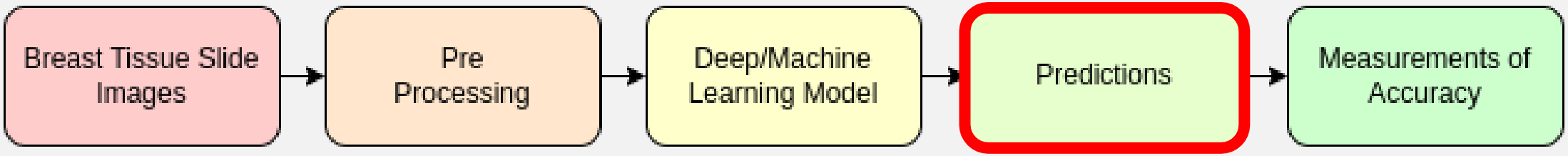
PC1	PC2	PC4	X
PC3	PC5	X	X
PC6	X	X	X
X	X	X	X

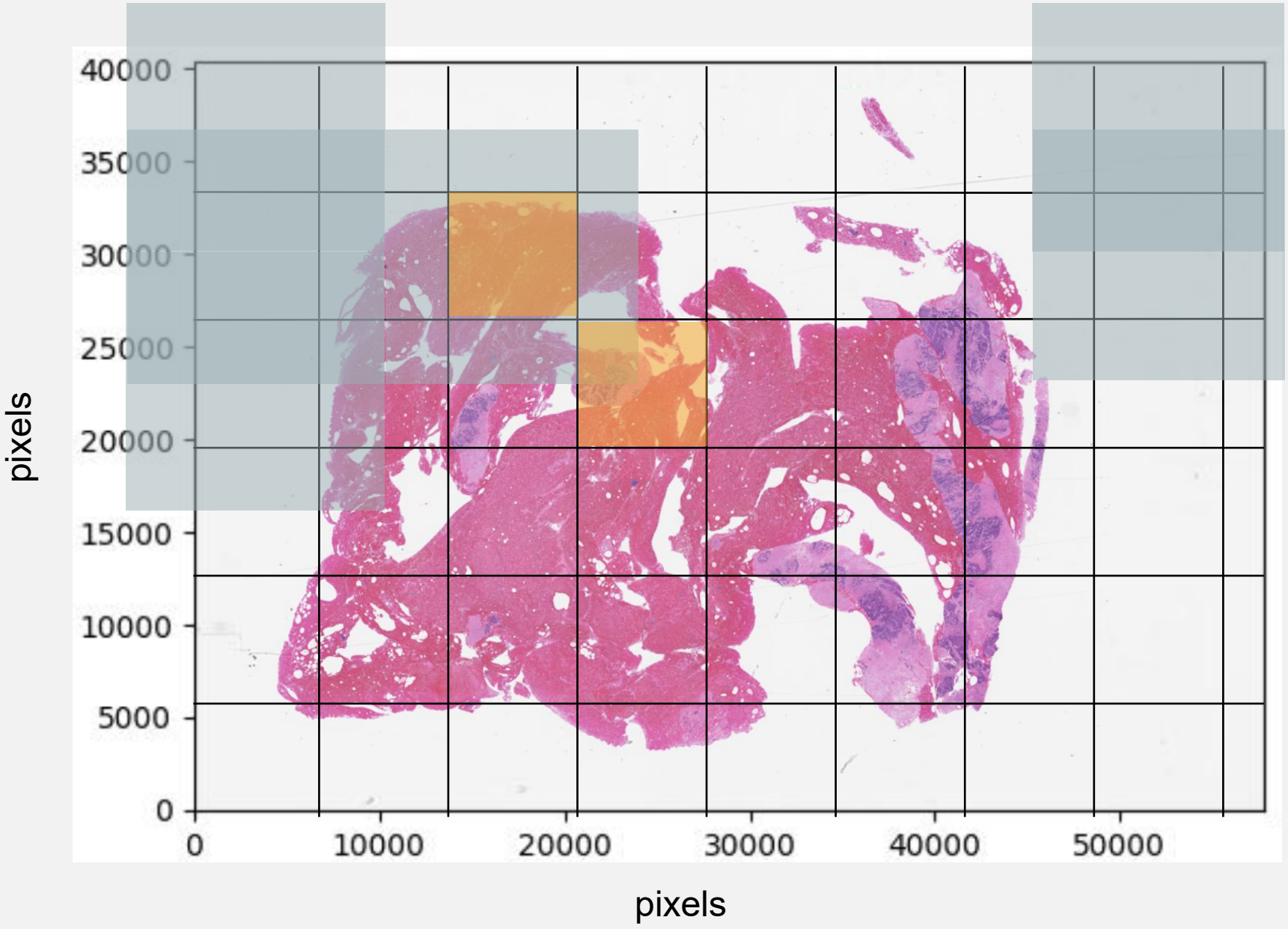
PRELIMINARY DESIGN

CONVOLUTIONAL NEURAL NETWORK

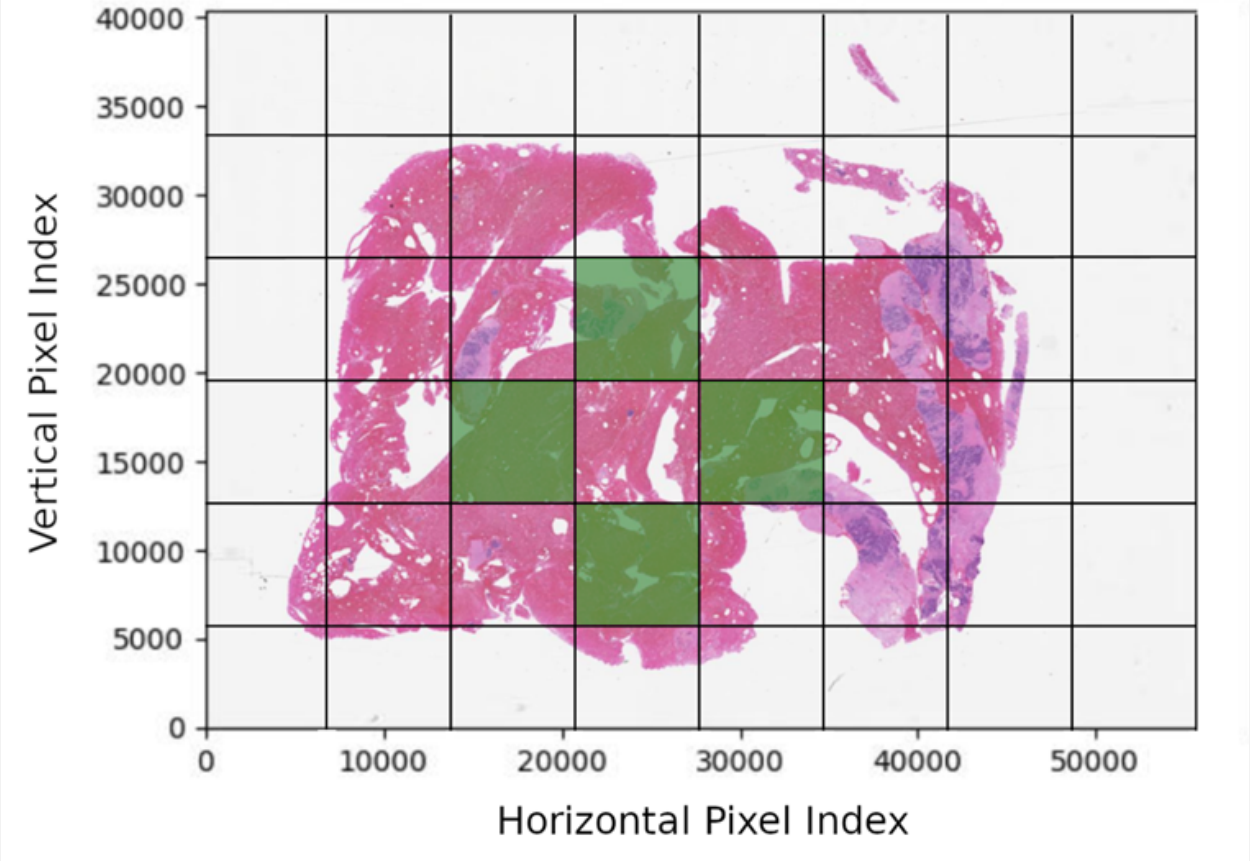


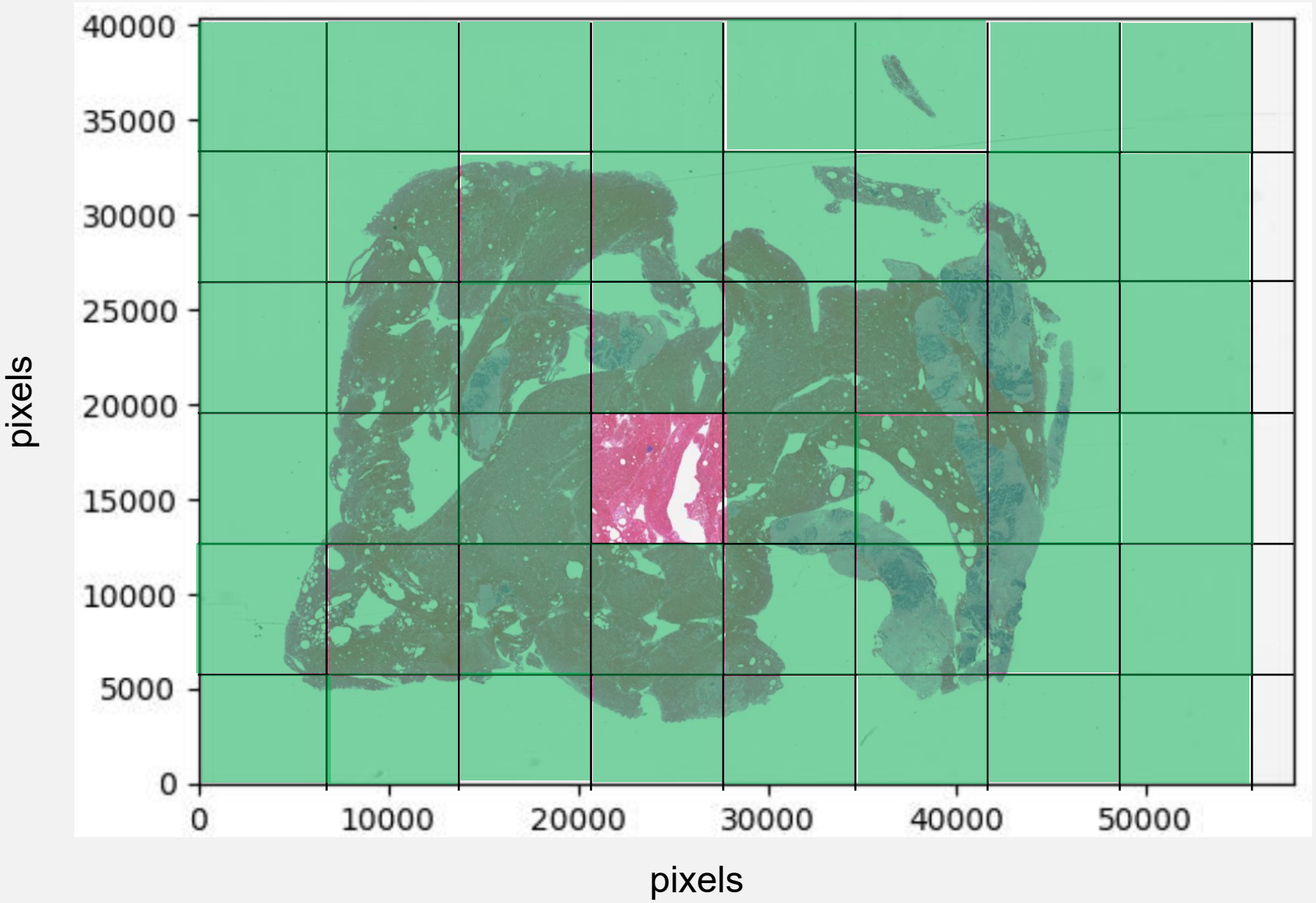
PRELIMINARY DESIGN



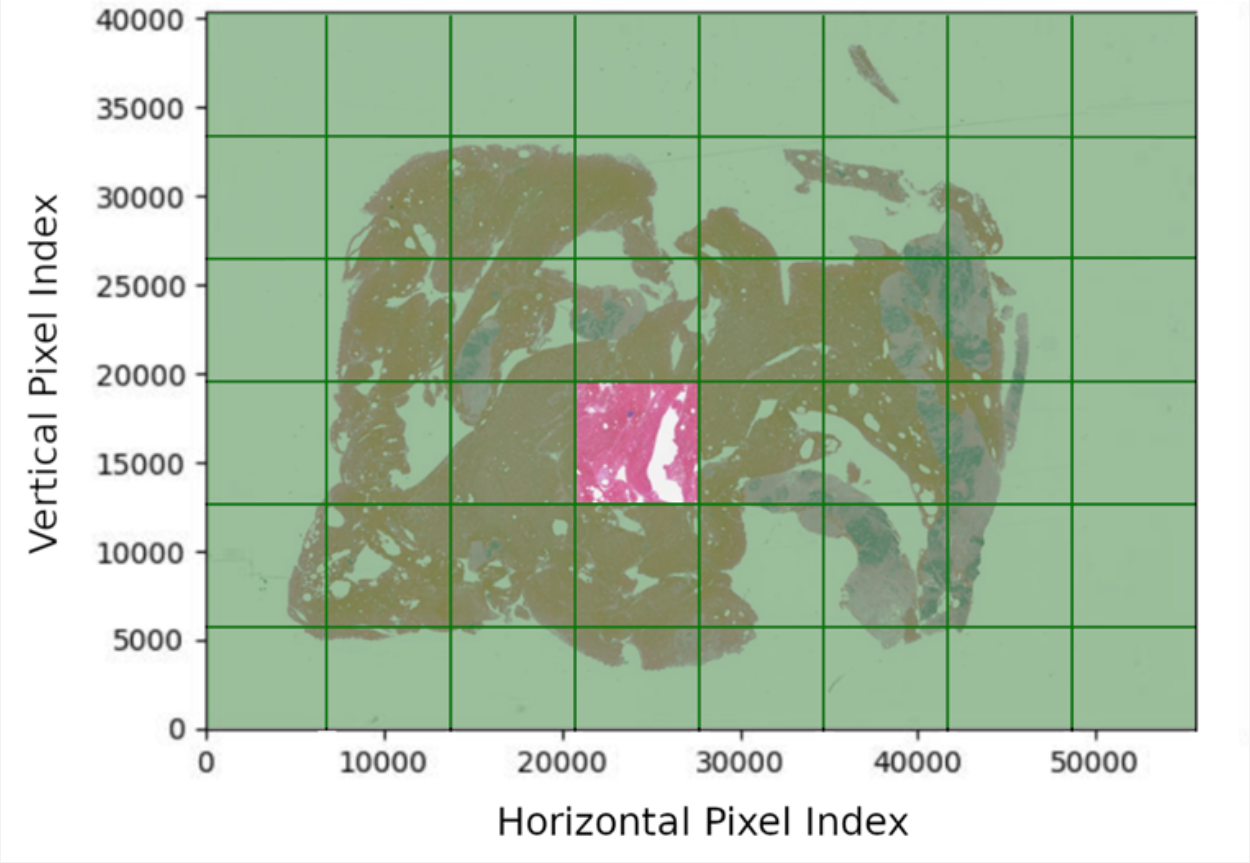

40000
35000
30000
25000
20000
15000
10000
5000
0
pixels
0
10000
20000
30000
40000
50000
pixels

FRAME TO PATCH

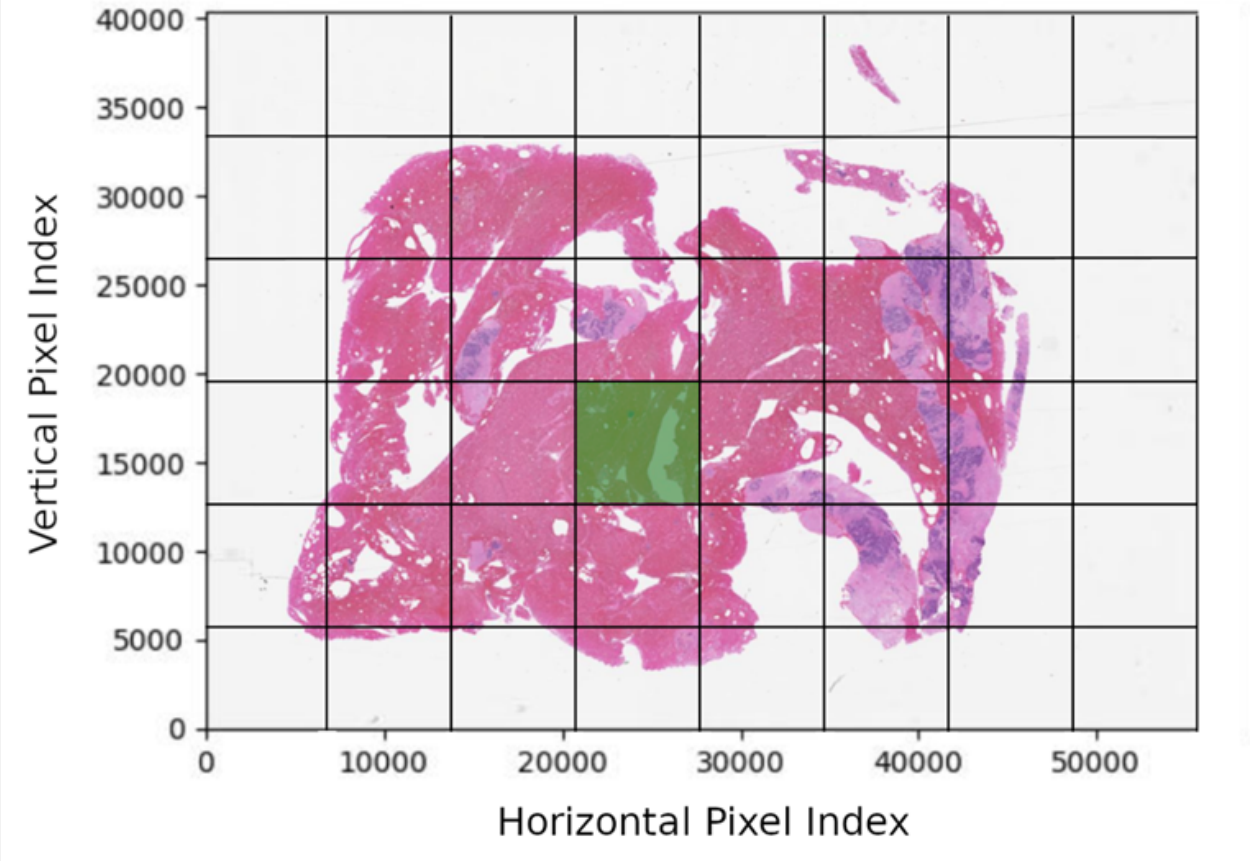



40000
35000
30000
25000
20000
15000
10000
5000
0
pixels
0
10000
20000
30000
40000
50000
pixels

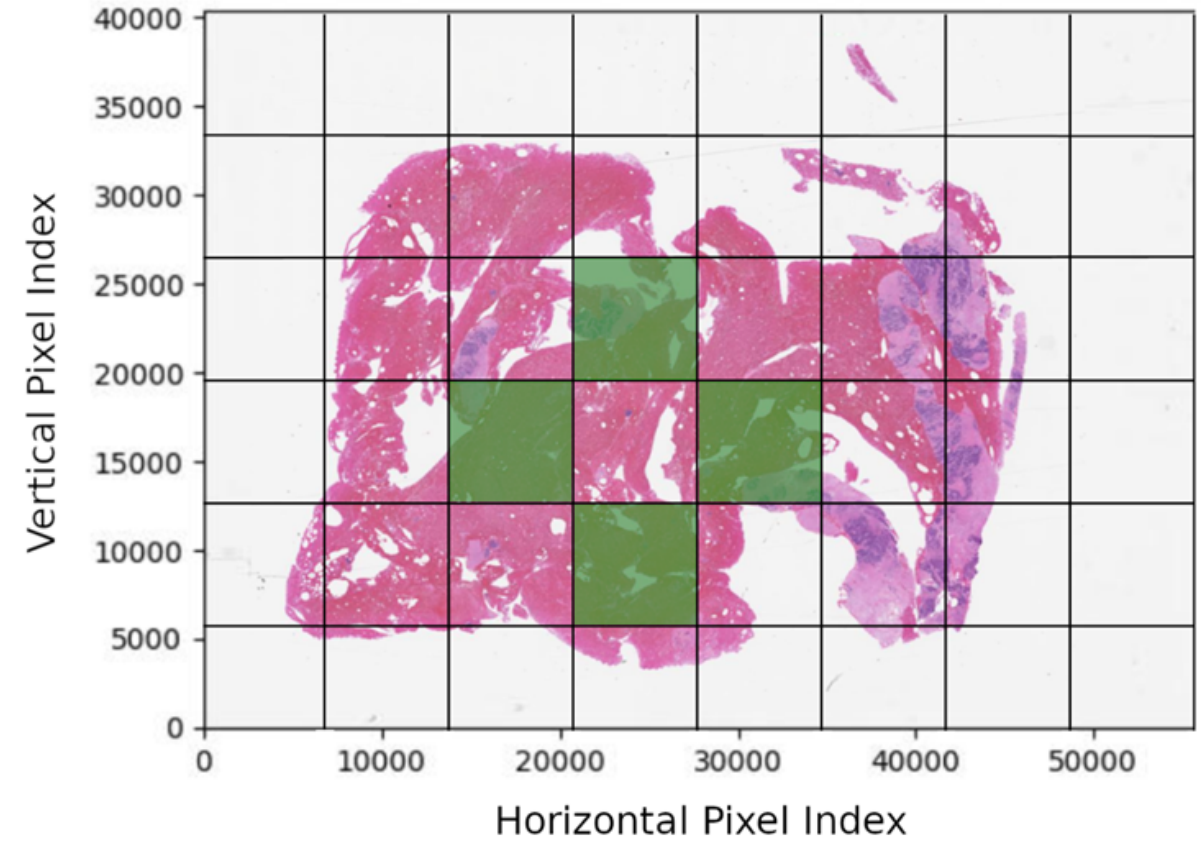
WINDOW TO PATCH

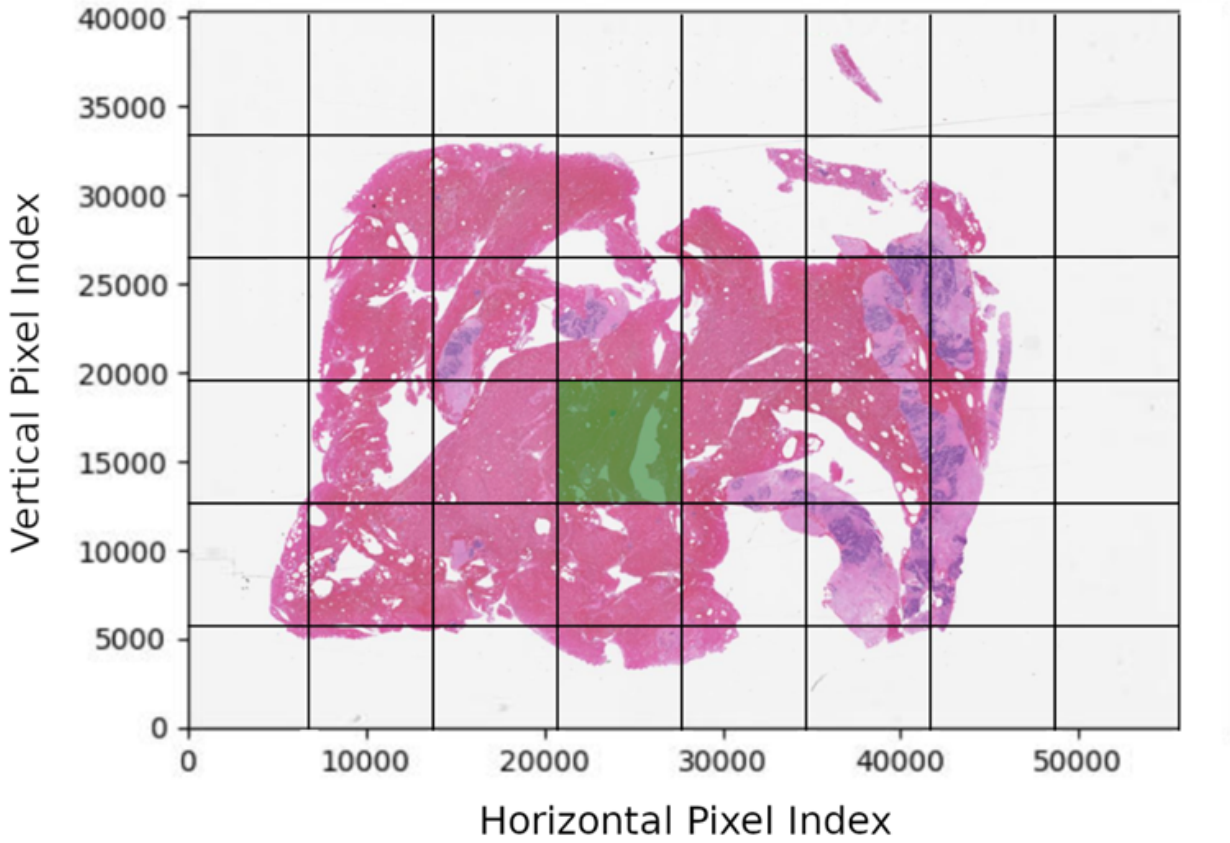


FRAME TO PATCH

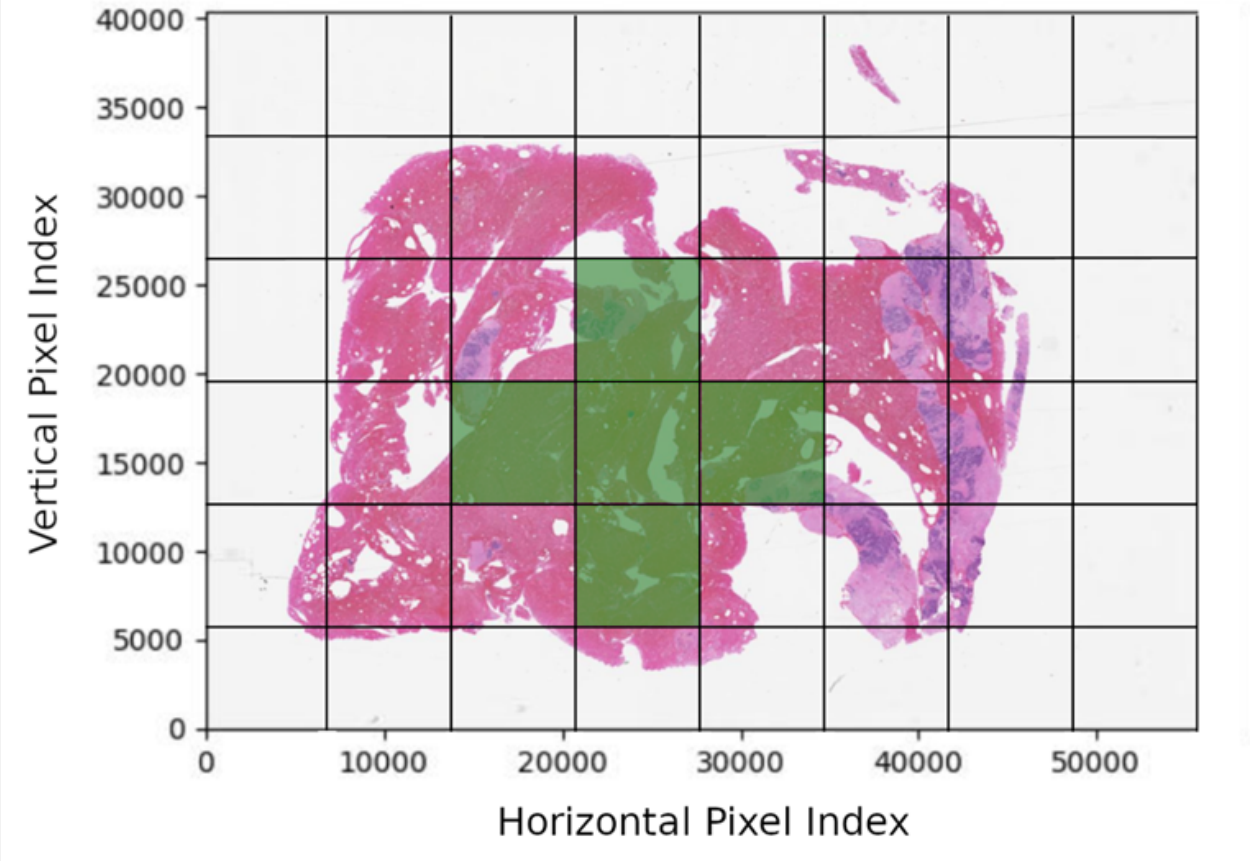


FRAME TO PATCH



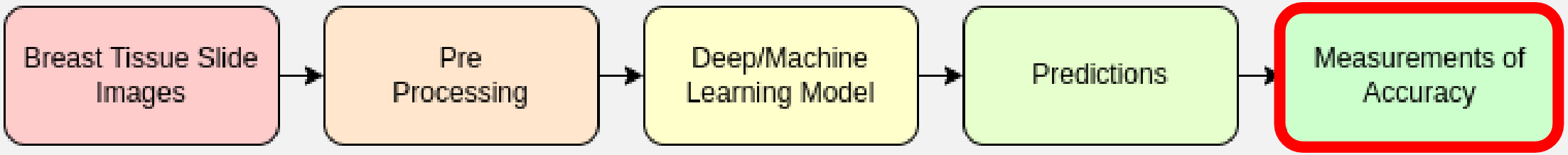


FRAME TO PATCH

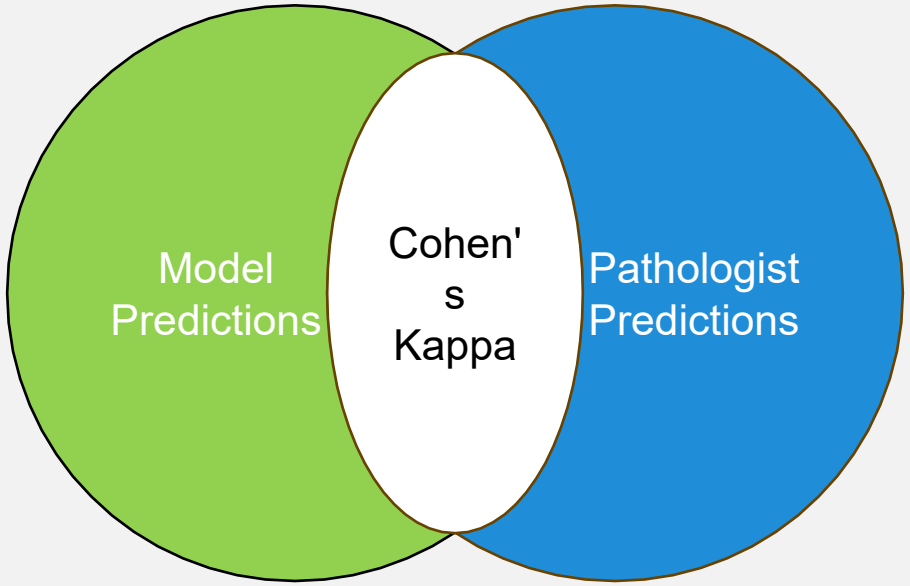


PATCH TO IMAGE LEVEL

PRELIMINARY DESIGN



FUTURE WORK



Cohen's Kappa > 60%
Source

F1 Score

$$\frac{TP}{TP + .5(FN)(FP)}$$

F1 Score > 90%
Source

Image Classification

No Sobel/Laplacian Filters

QUESTIONS?

Acknowledgements:
Dr. Joseph Picone – Mentor & Data Coordinator
Claudia Dumitrescu – AI Expert
Phuykong Meng – GUI Planning & Scoring

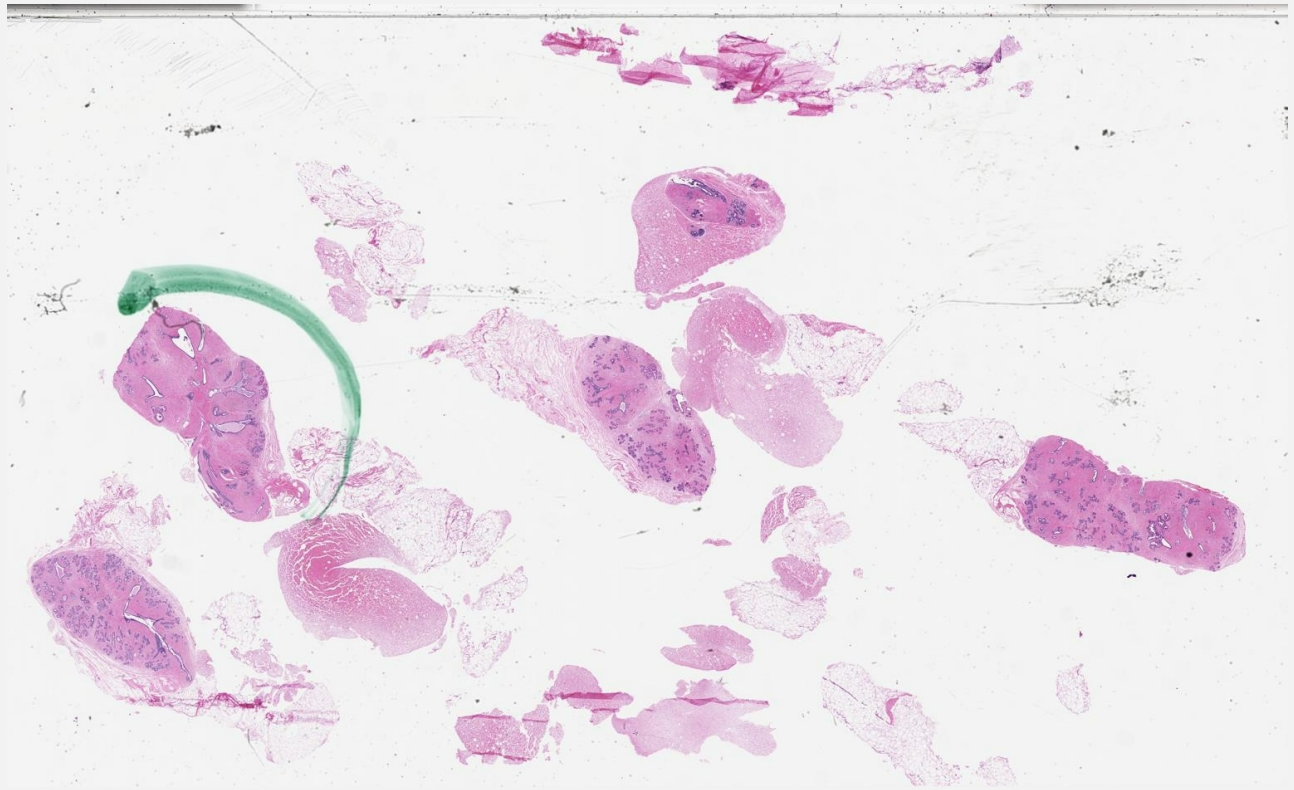
For the curious

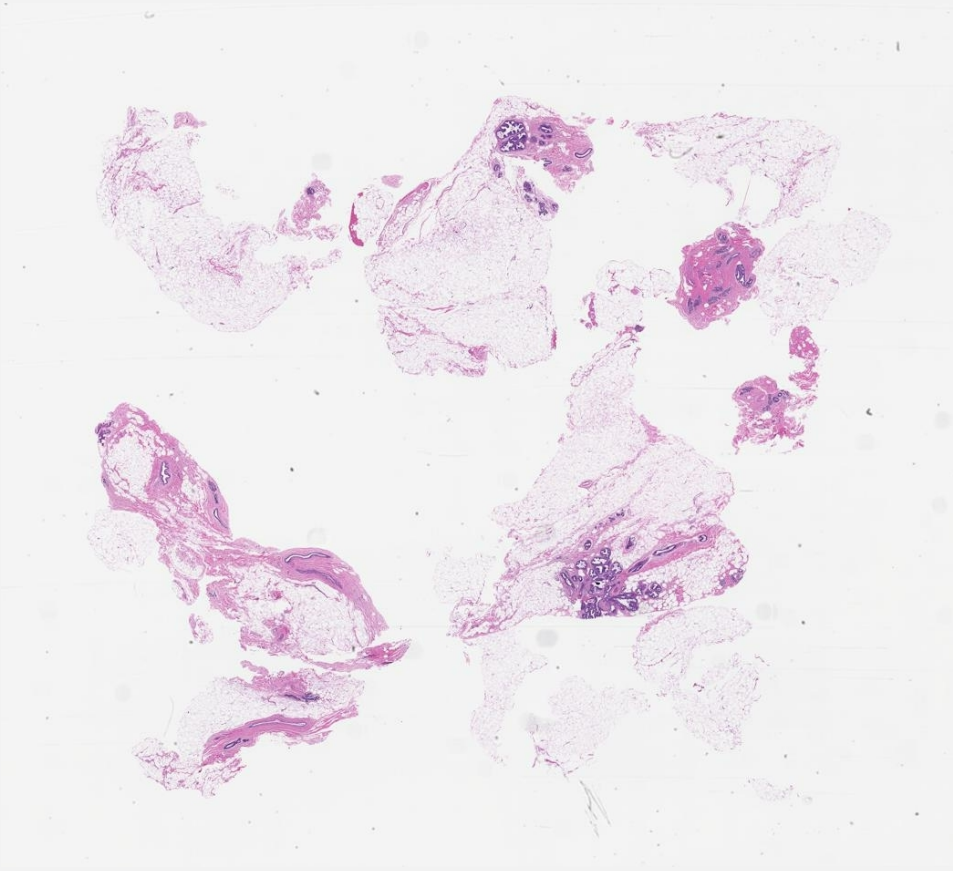
- *Benefits of Machine Learning in Healthcare*
- *Machine Learning in Healthcare*
- *What is Machine Learning in Healthcare?*
- *Significance of Machine Learning in Healthcare*
- *The Potential for Artificial Intelligence in Healthcare*

PROGRESS SINCE LAST PRESENTATION

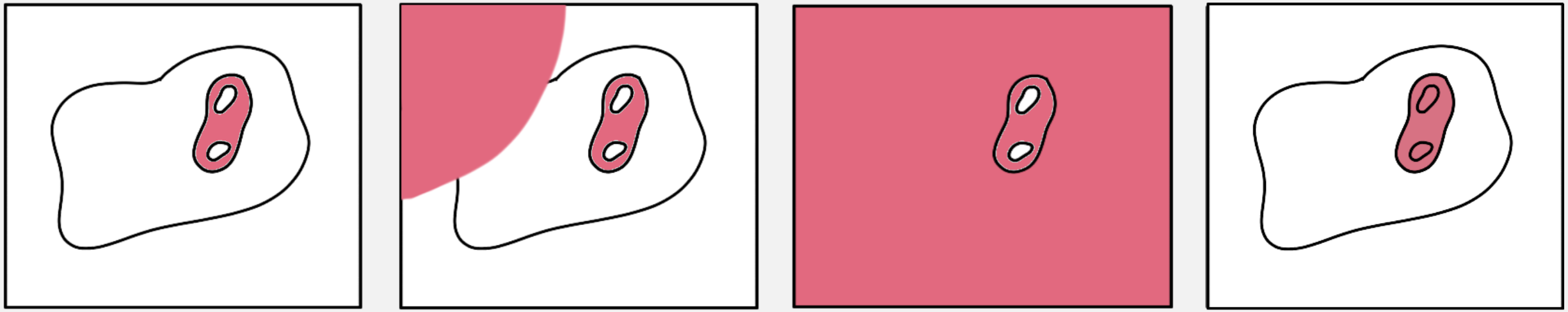
- 2D Discrete Cosine Transform (DCT)
- Principal Component Analysis (PCA)
- Convolutional Neural Network (CNN)

TUHDP BIOPSY SLIDE SAMPLE

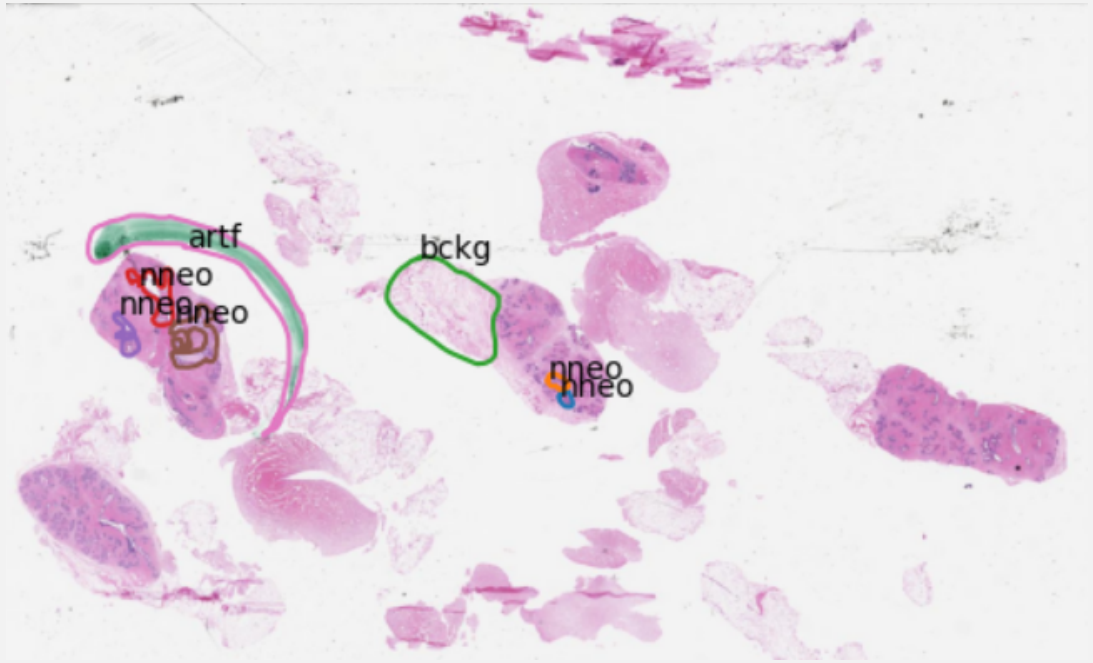


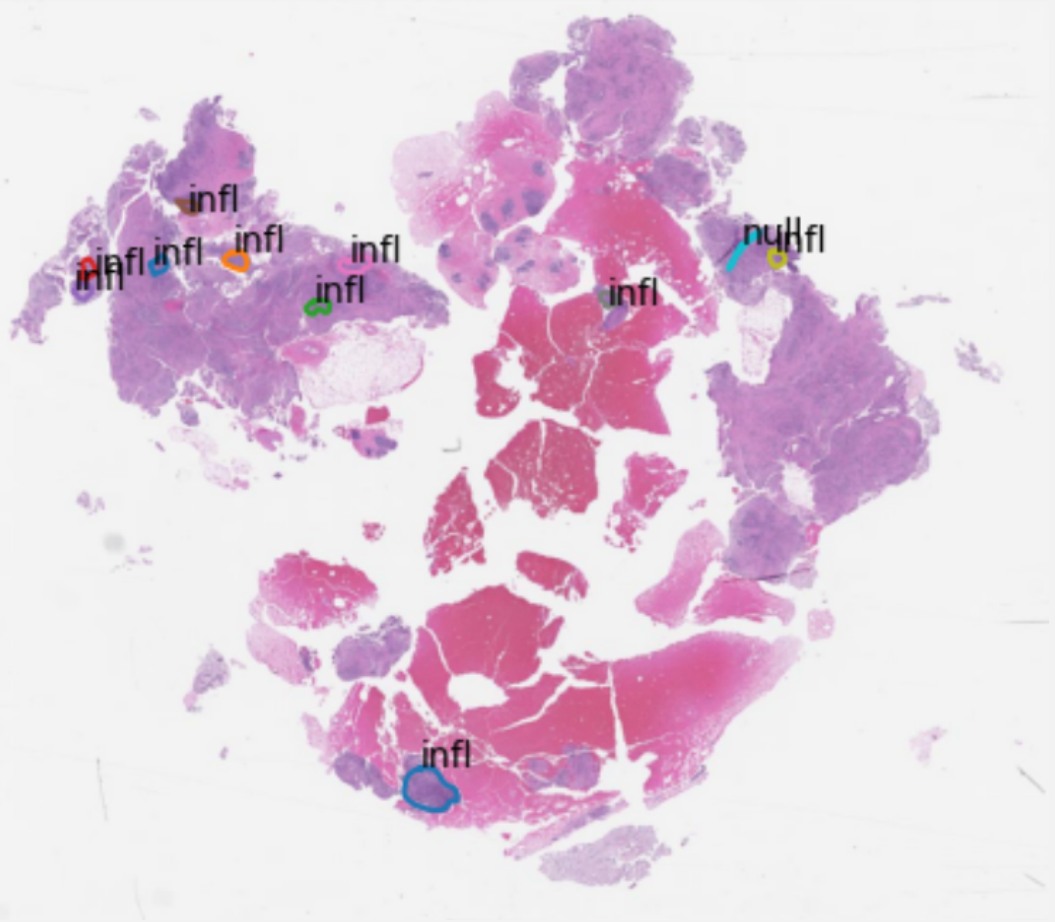


WINDOW TO PATCH

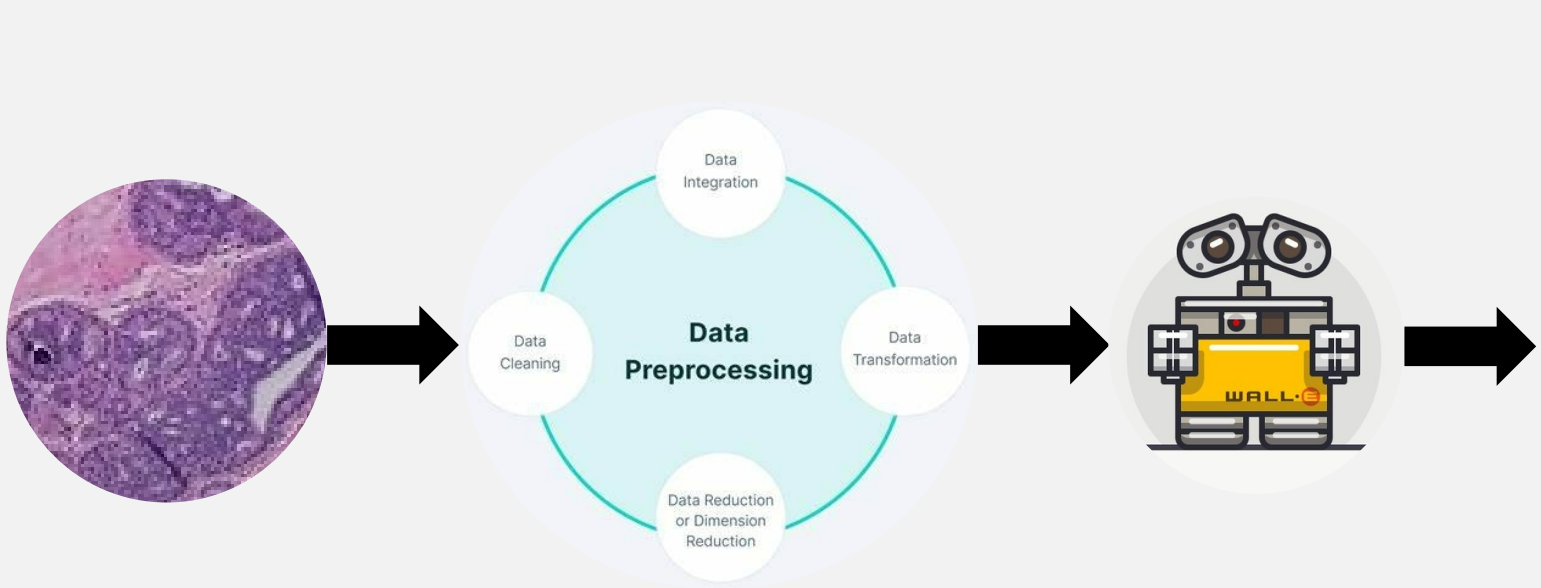


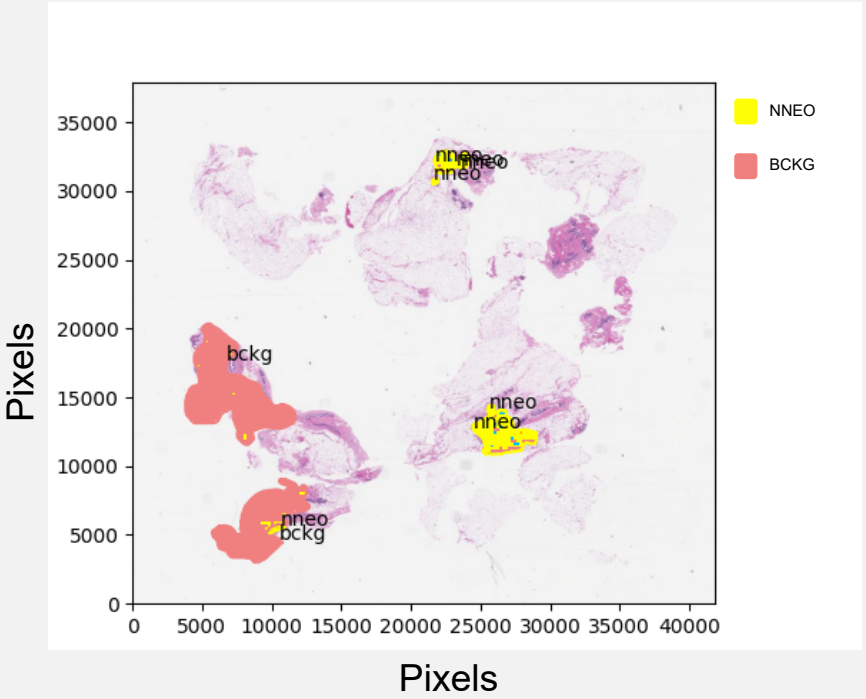
TUHDP BIOPSY SLIDE SAMPLE



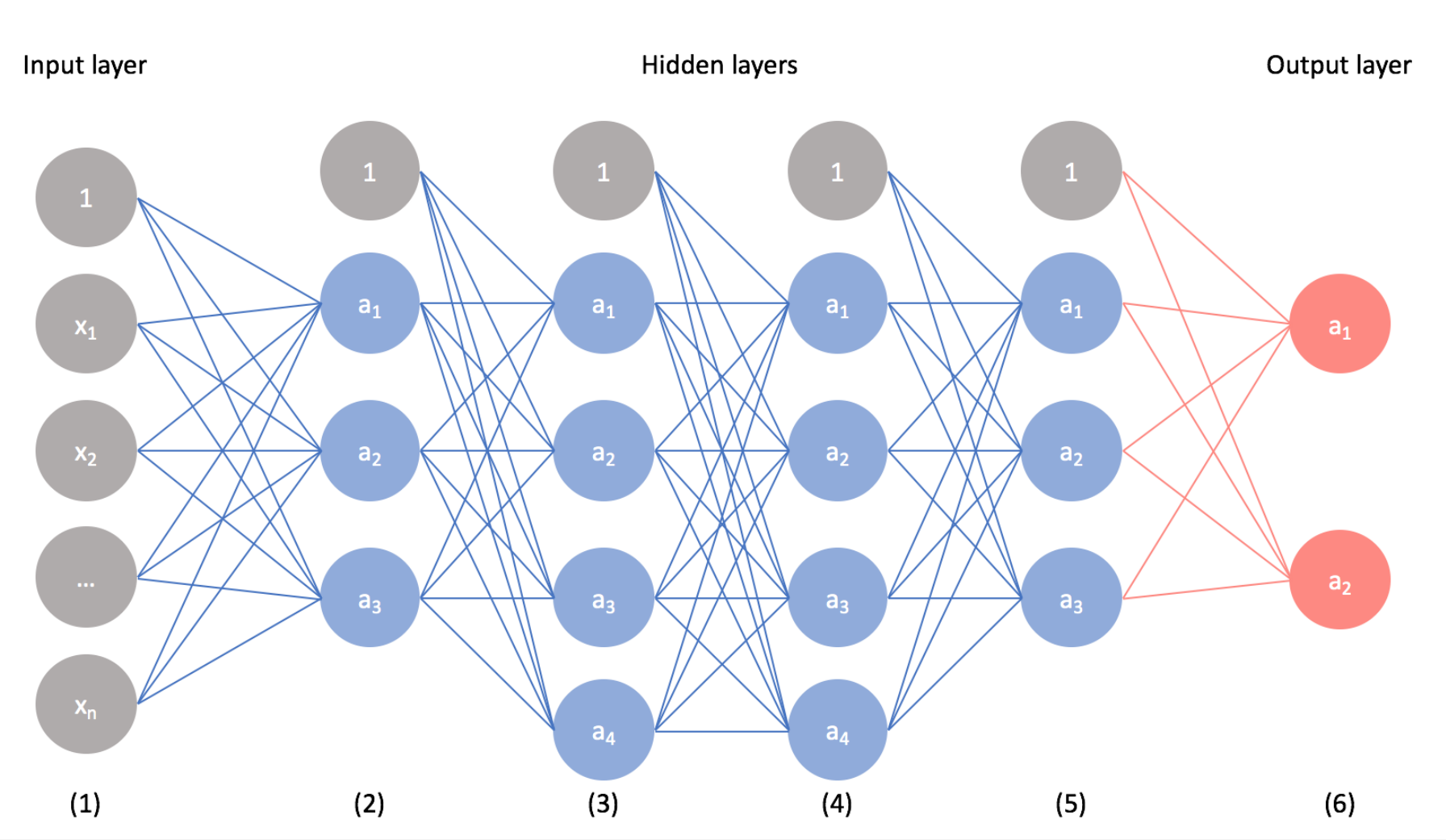


DEEPER LOOK AT MODEL I/O





CONVOLUTIONAL NEURAL NETWORK

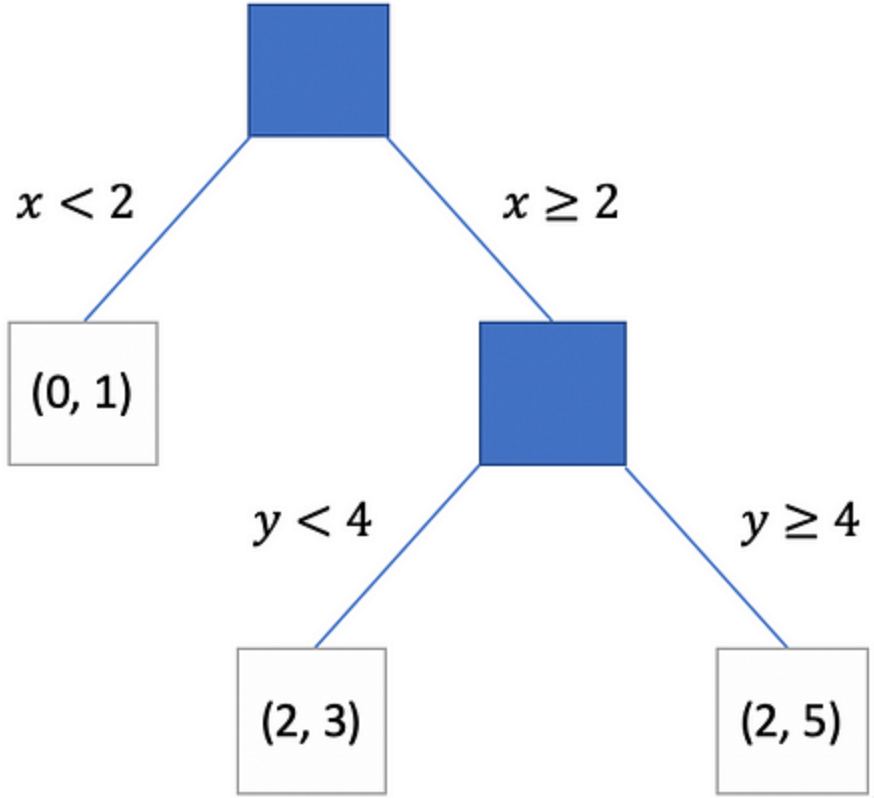


RANDOM FOREST

Dataset

x	y
0	1
2	3
2	5

Isolation Tree

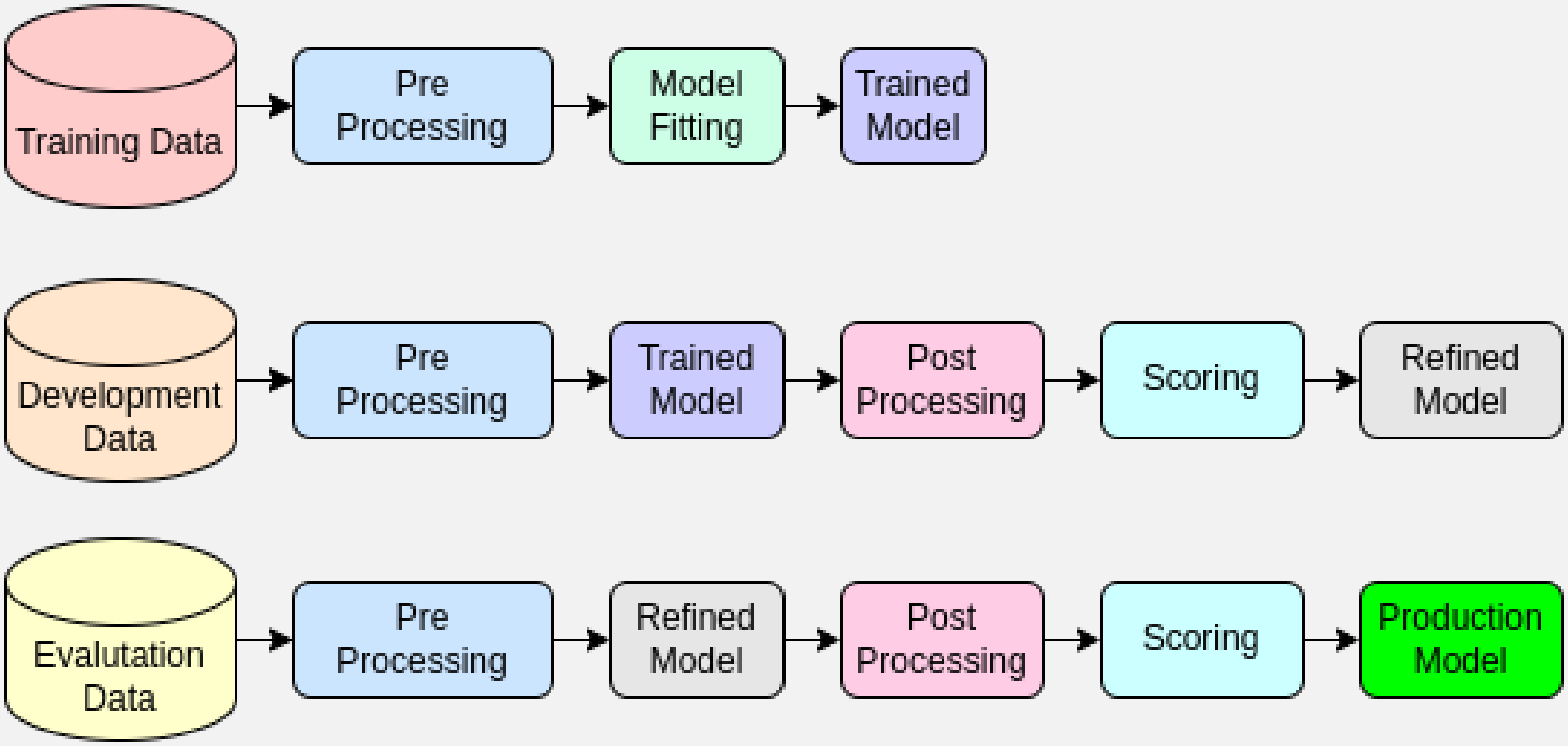


WHAT IS A MODEL?

- Three essential components for machine learning
  - Data
  - Model
  - Training algorithm
- Model
  - Mathematical representation of relationships in the data
- Training algorithm
  - Uses data
  - Adjusts variables *in the model* until output matches input

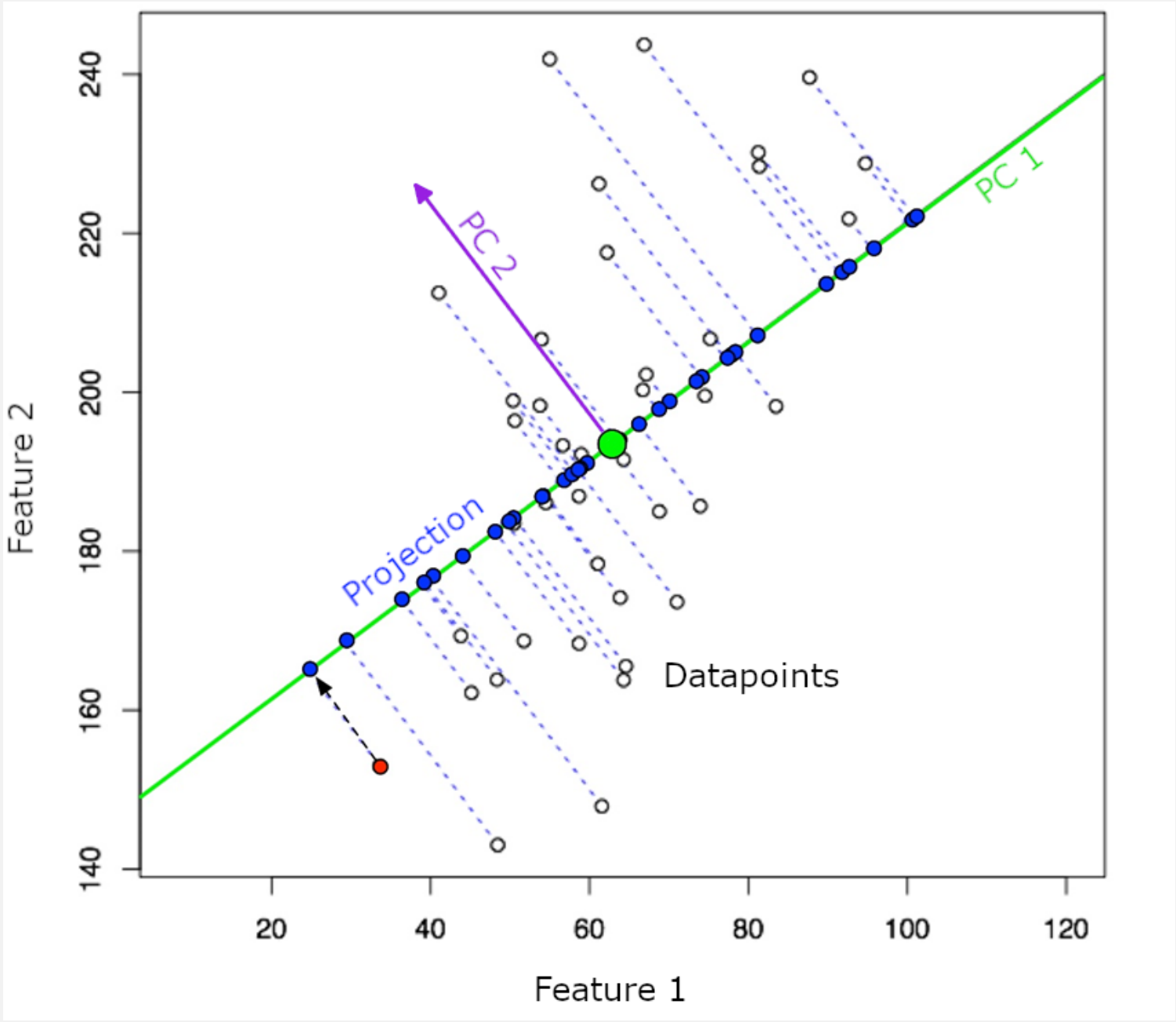
TEMPLE UNIVERSITY HEALTH DIGITAL PATHOLOGY CORPUS

3,505 Tissue Images 1.23 Terabytes



PRINCIPAL COMPONENT ANALYSIS (DIMENSION REDUCTION)

- *After* feature generation
- Also belongs in the pre-processor stage after digitizing/segmenting slides and before training
- Reduces the number of features while minimizing information lost from feature reduction
- Fewer features compared to no PCA or dimension reduction


PC 1
PC 2
Projection
Datapoints
Feature 1
Feature 2
20
40
60
80
100
120
140
160
180
200
220
240

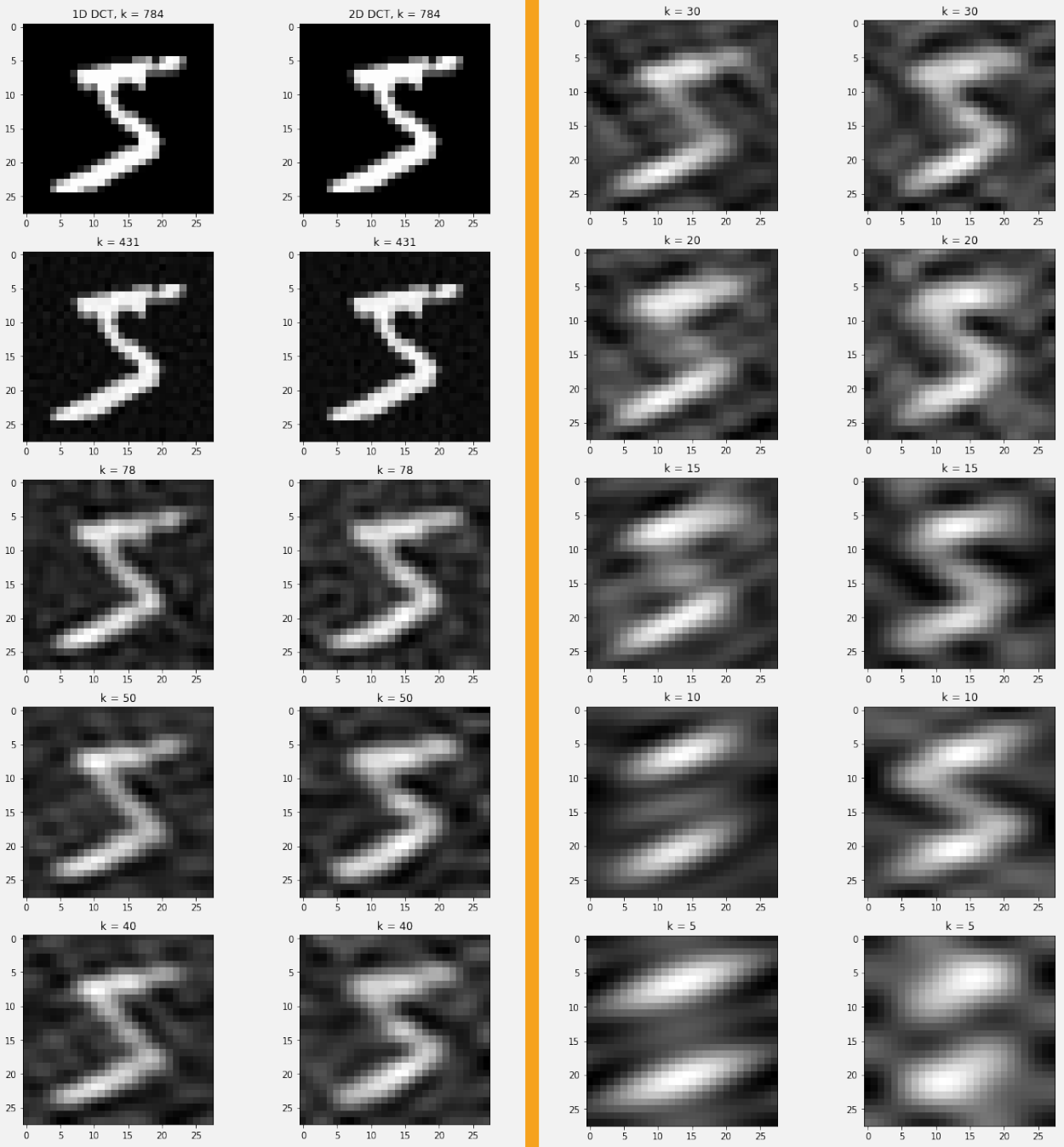
TESTING GOALS

2-DIMENSIONAL DCT

- Part of feature generation
- Belongs in the pre-processor stage after digitizing/segmenting slides and before training
- Greater spectral/energy density than one-dimensional DCT
- Therefore, fewer features compared to one-dimensional DCT

2-DIMENSIONAL DCT

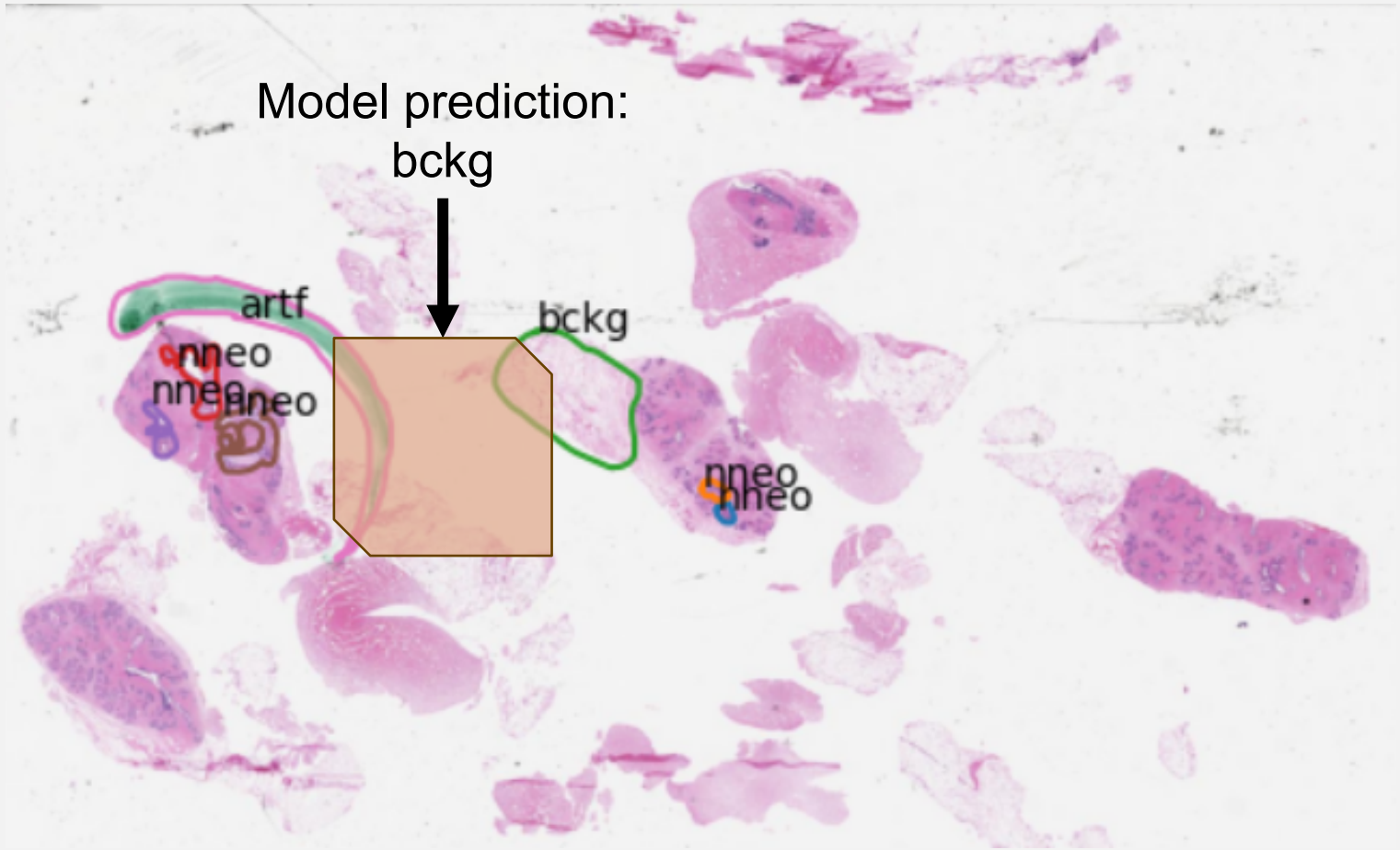
- Belongs in the pre-processor stage
- Converts colors to frequencies
- Allows us to retain fewer features



SCORING

- Find overlapping area between the model's predictions and the human pathologist's predictions
- Use overlapping area to construct a confusion matrix
- Use confusion matrix to generate F_1 scores and Cohen's kappa coefficients

OVERLAPPING AREA



		Pathologist (True)										
	Frame Labels	Unlab	Bckg	Norm	Null	Artf	Nneo	Infl	Susp	Indc	Dcis	Sums across rows
Model (Predicted)	Unlab	20	1	4	1	7	6	7	0	4	6	56
	Bckg	6	30	1	3	5	2	4	7	3	3	64
	Norm	3	2	29	6	0	1	1	3	1	2	48
	Null	3	0	4	24	3	4	6	3	5	0	52
	Artf	5	6	4	7	29	3	6	2	3	3	68
	Nneo	1	6	6	1	3	21	5	3	3	5	54
	Infl	6	3	7	4	0	6	26	7	1	2	62
	Susp	6	4	6	4	5	3	6	22	0	6	62
	Indc	5	5	2	6	0	3	6	0	28	0	55
	Dcis	5	0	6	6	7	7	4	7	5	24	71
	Sums across columns	60	57	69	62	59	56	71	54	53	51	592

		Pathologist (True)		
	Frame Labels	NNeo	Not NNeo	Sums across rows
Model (Predicted)	NNeo	21	33	TP+FP = 54
	Not NNeo	35	503	FN+TN = 538
	Sums across columns	TP+FN = 56	FP+TN = 536	592

FILTERS

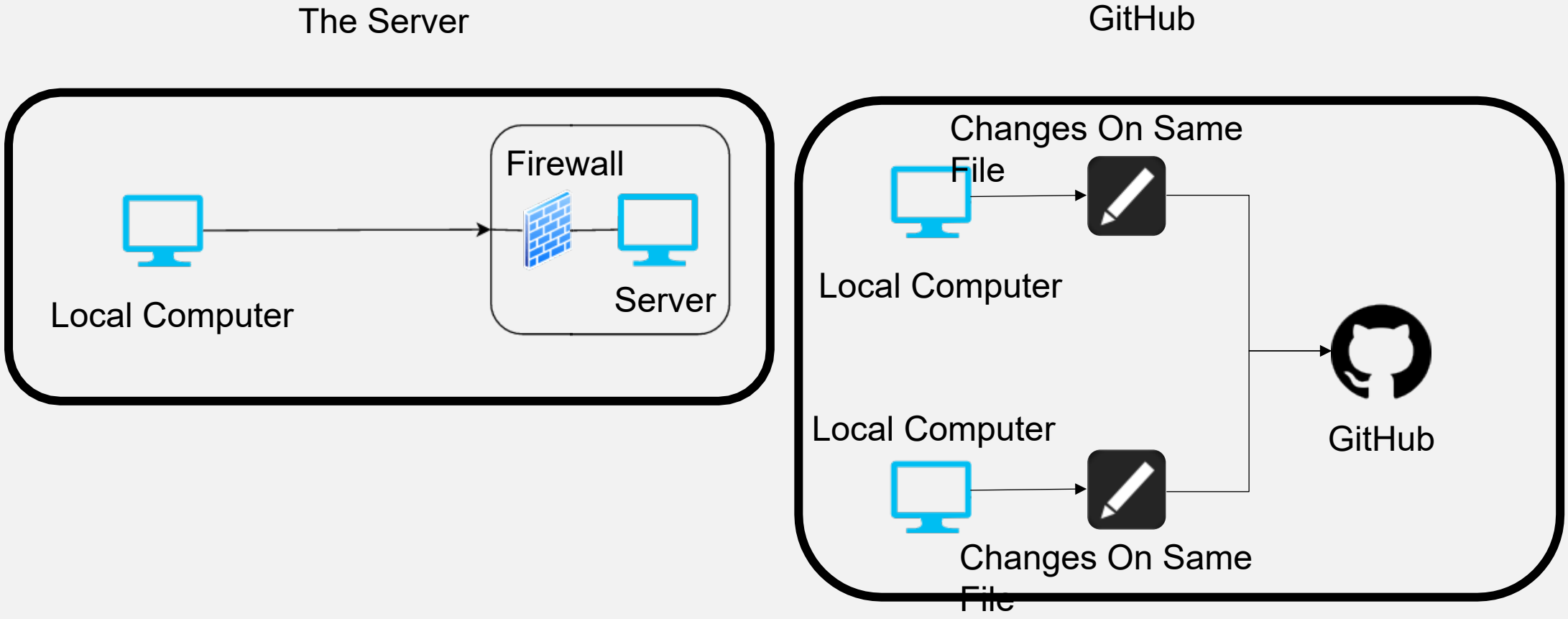
Original image

Filter “Laplace” applied

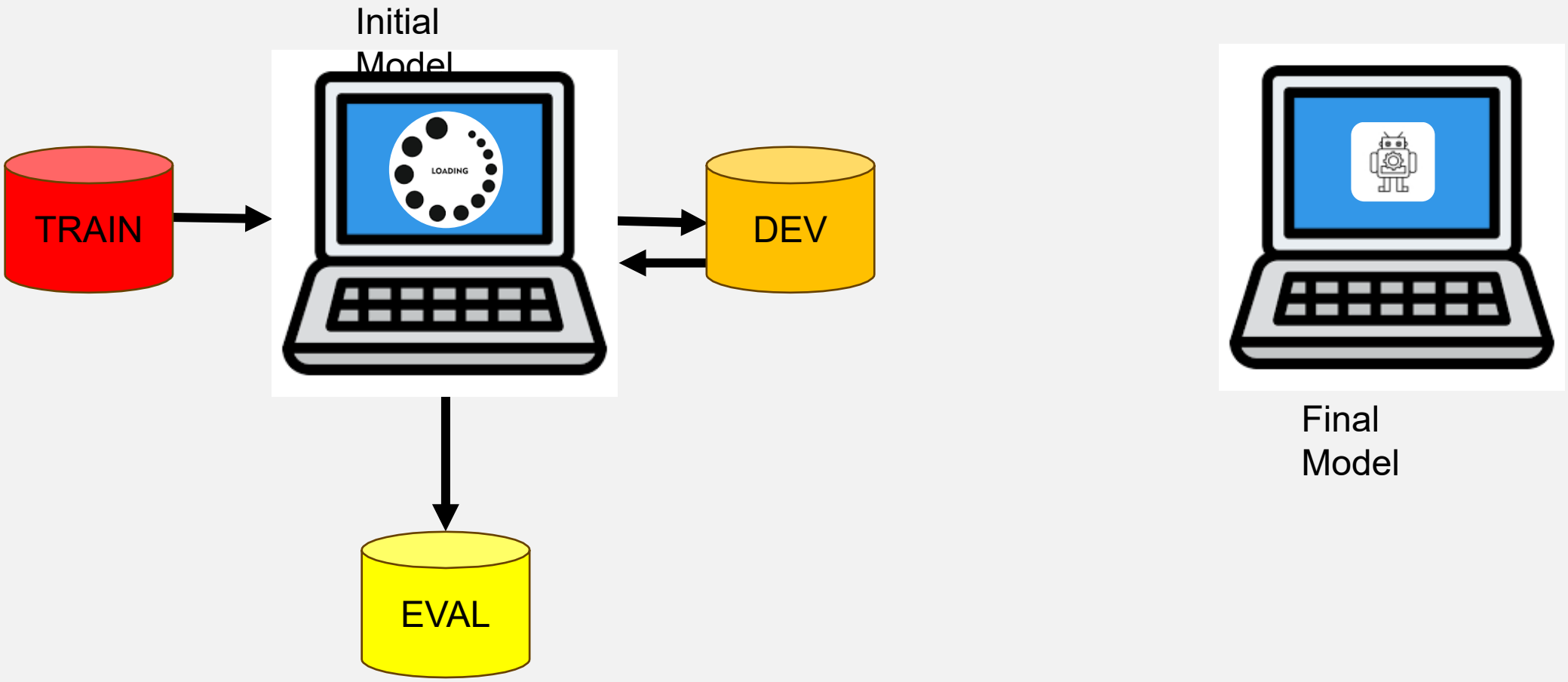
FRAME LEVEL EVALUATION RESULTS

	Dataset	Accuracy Rate [%]
Random Forest	TRAIN	100.00
	DEV	86.33
	EVAL	85.87

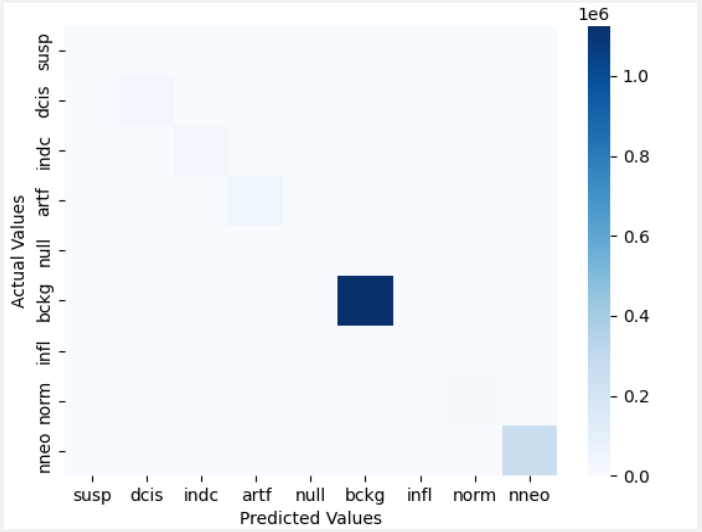
GITHUB AND THE SERVER



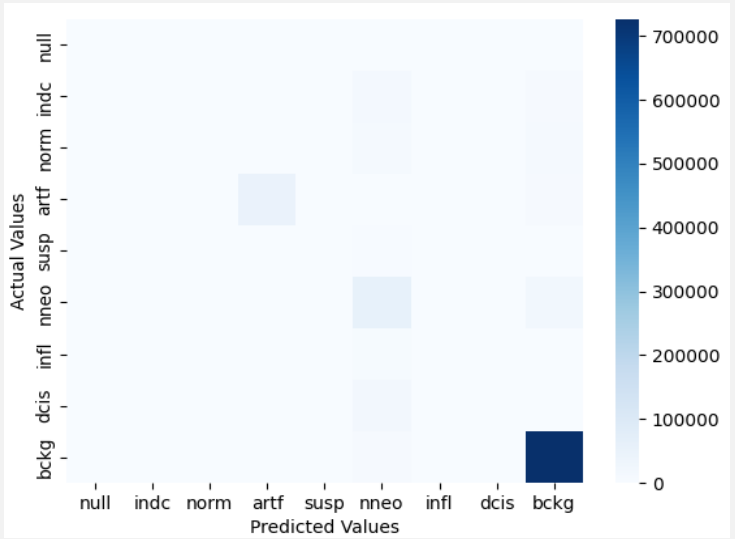
EXPLAIN TRAIN, DEV, AND EVAL IN THE PROCESS



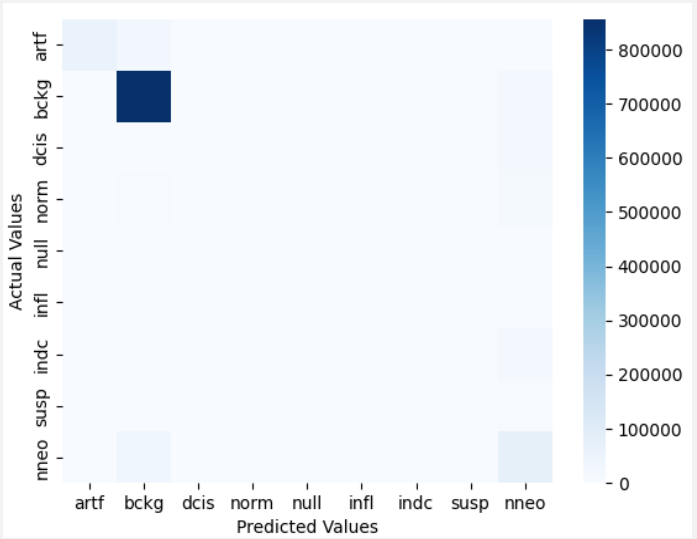
RNF CONFUSION MATRIX



TRAIN

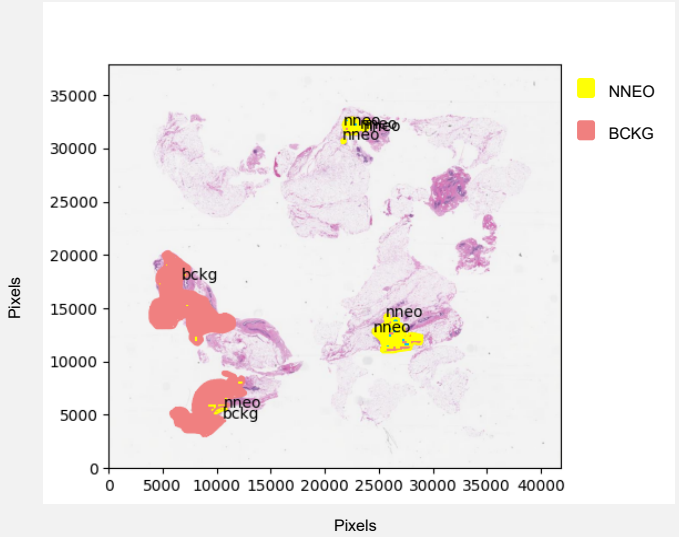


DEV



EVAL

RNF DECISION SURFACES



HOW THIS FITS