Breast Cancer Detection with Computer Vision

Arjun Rao, Emily Wilkins, Javeria Rangoonwala,
Skyler Saleebyan, Vikrant Vaidya

Why Breast Cancer ?

1/8 women are expected to be diagnosed with cancer over their lifetime

40+ women are recommended to undergo annual screening

100x more likely in women than men

~300,000 people are expected to be diagnosed in the US each year

Presentation Overview

01
Introduction
The context of the problem

02
The Dataset
Data source and format

03
Preprocessing
Preprocessing steps on the data

04
Methodology and Analysis
Our methods and results

05
Future Steps
What we plan to do next

Data Set

- Provided by ICIAR 2018 competition
- Breast tissue microscopy images stained with Hematoxylin and Eosin (H&E)
- For data preparation, the slides were verified by multiple histologists and tossed if there was disagreement

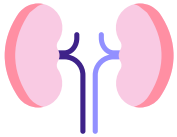
Histology is the tipping point

Palpation/mammogram suspects a mass

Histology confirms/denies presence of cancer

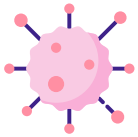
Cancer treatment begins:
- Mastectomy
- Chemotherapy
- Radiation Treatment
- Hormone Therapy
- Targeted Therapy

Misdiagnosis



$20-100k

8% Misdiagnosis

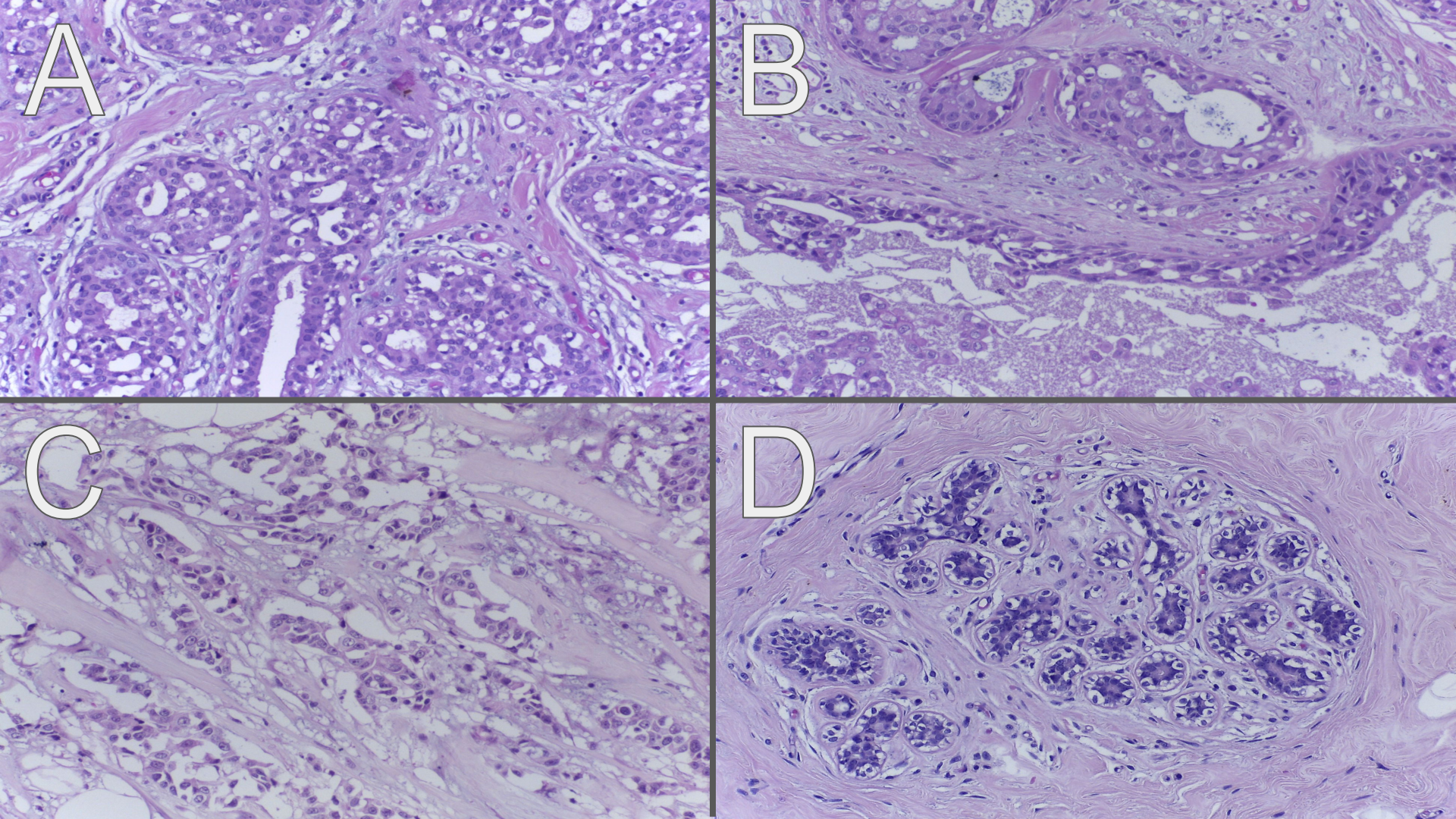


30% Discrepancies

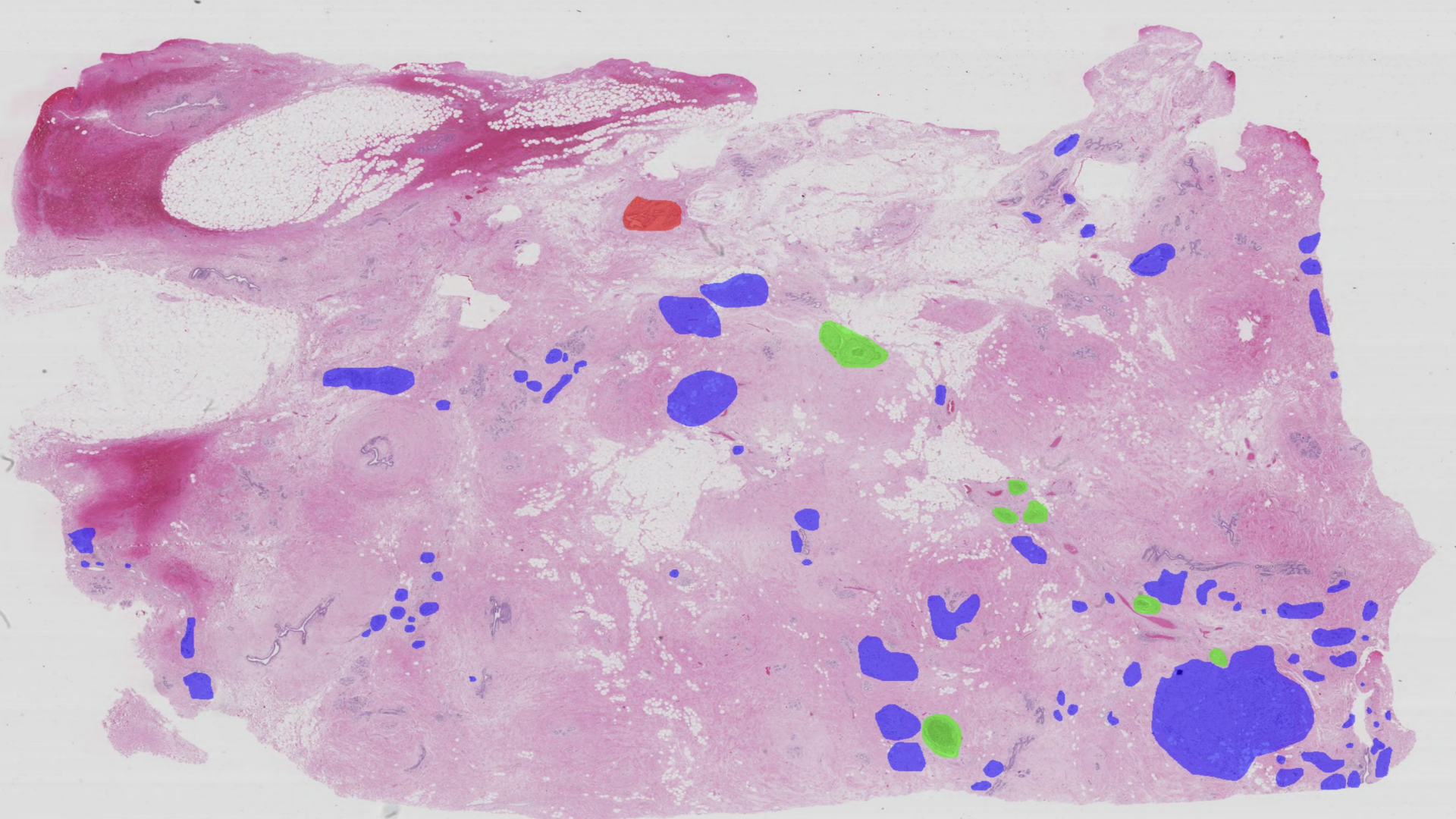
Lax Industry Standards

Early Stage Detection

Spot 5 differences:


A
B
C
D

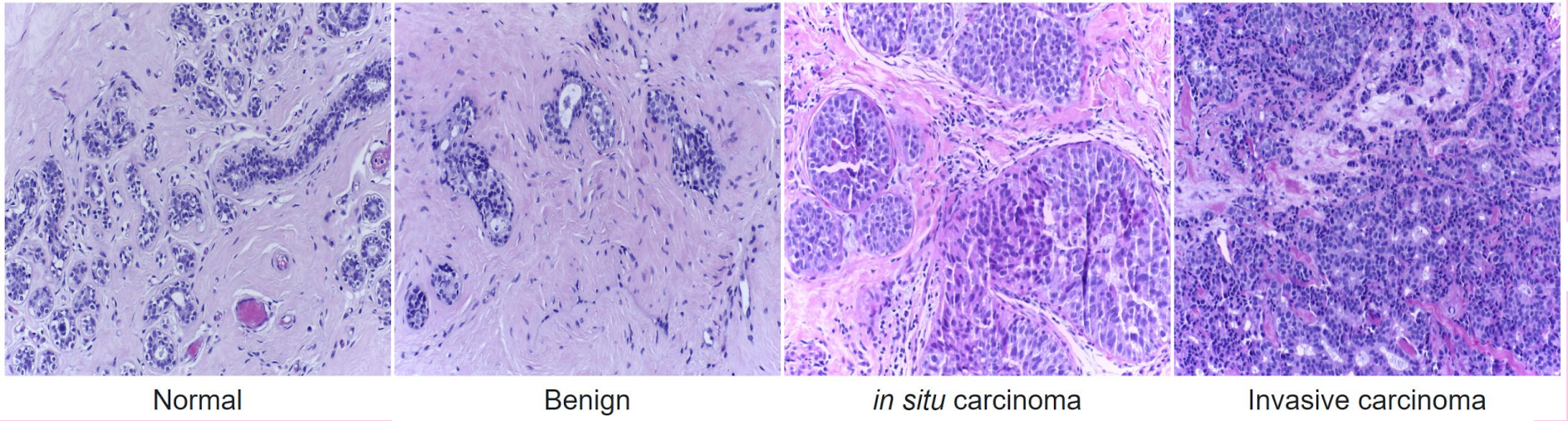

HUISKENS
'20



Microscopy Images

There are 4 types of images:

- **Normal: 100**
- **Benign: 100**
- **In situ carcinoma: 100**
- **Invasive carcinoma: 100**



What do we Hope to Achieve

SHORT TERM

Understand and implement -

A. Transfer learning - ResNet, VGGNet, Inception

B. State of the art ensemble techniques - LightGBM, XGBoost

LONG TERM

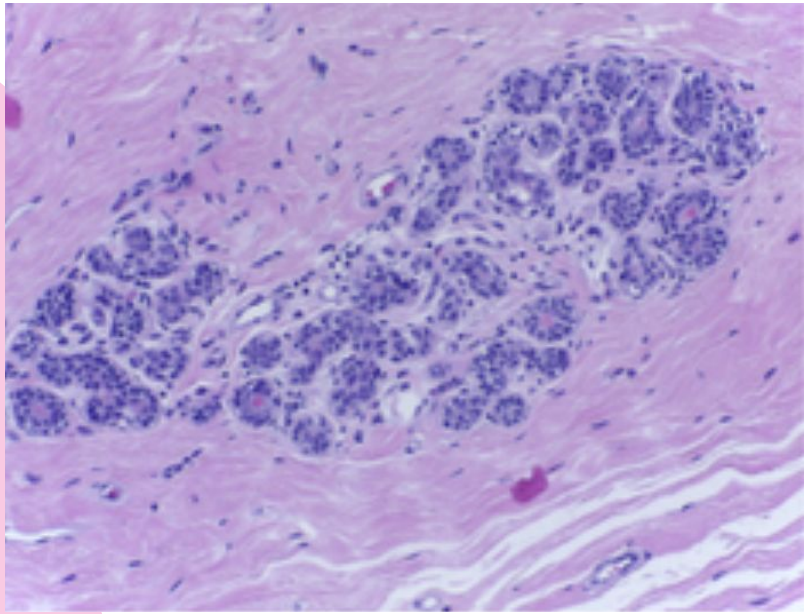
A. Incorporate Layer-wise Relevance Propagation (heat mapping) to identify important pixels

B. Fine tune / tweak

C. Interact with industry personnel for insights into current implementations

Into the Data: Preprocessing (1)

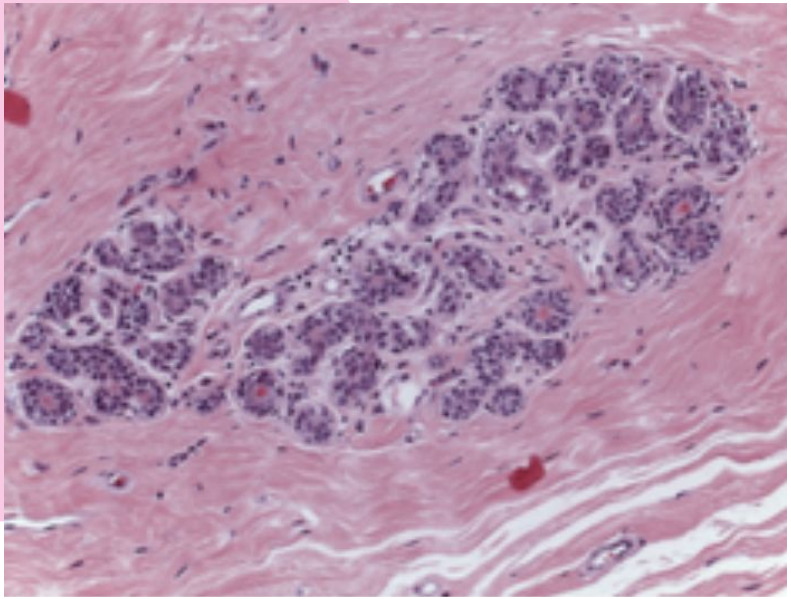
Original Image



- H&E stain images (2048 x 1536 pixels)
- All images are digitised
  - Magnification of 200
  - Pixel size of 0.42µm x 0.42µm

Into the Data: Preprocessing (2)

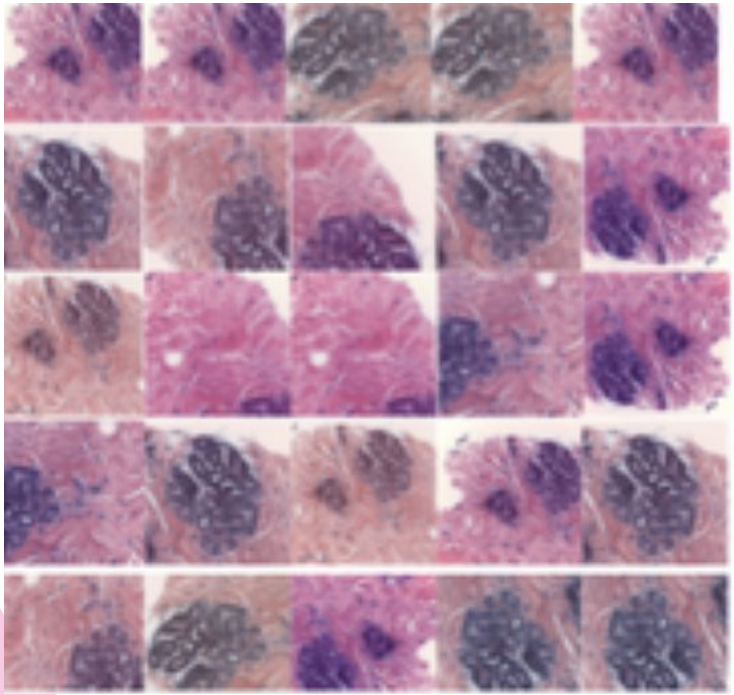
Normalization of Stain



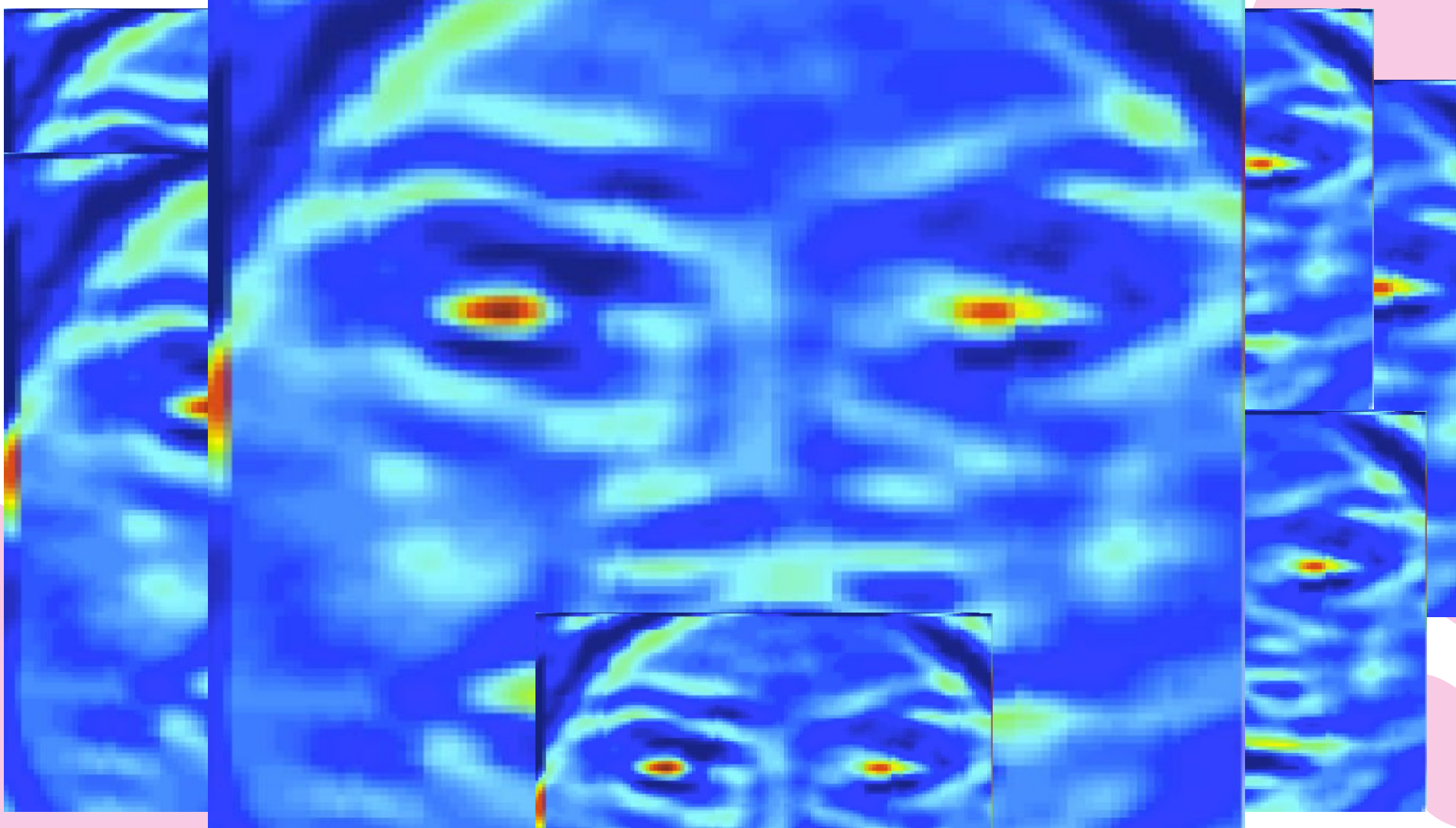
- ➔ Each image undergoes staining normalisation to maintain consistency in features
- ➔ A random H&E augmentation is performed b/w a high (1.3) and low(0.7) threshold
- ➔ Overall **50** random color augmentations per image

Into the Data: Preprocessing (3)

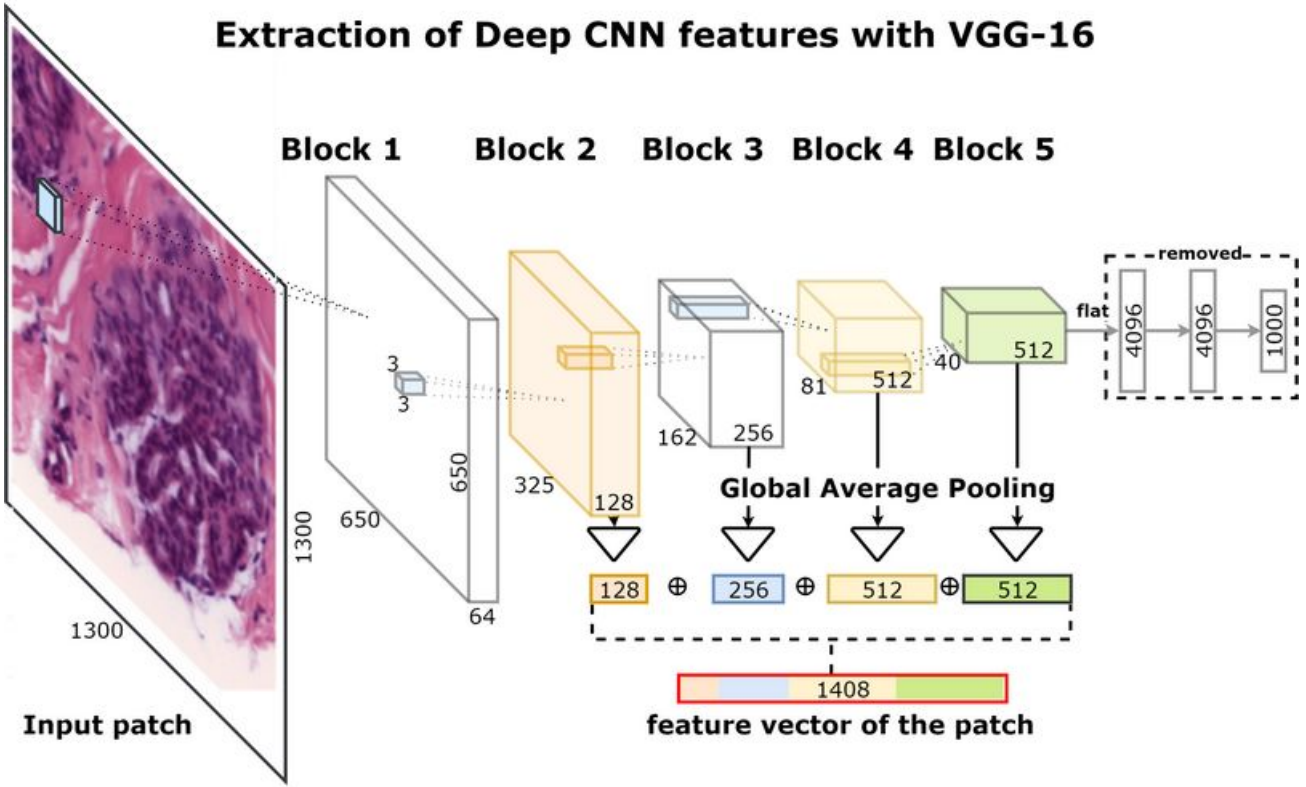
Augmented Crops



- **2** patch sizes chosen for normalised images **[400,650]**
- Each image undergoes random spatial zoom augmentations
- Resulting image used to extract **20** total crops with/without rotations



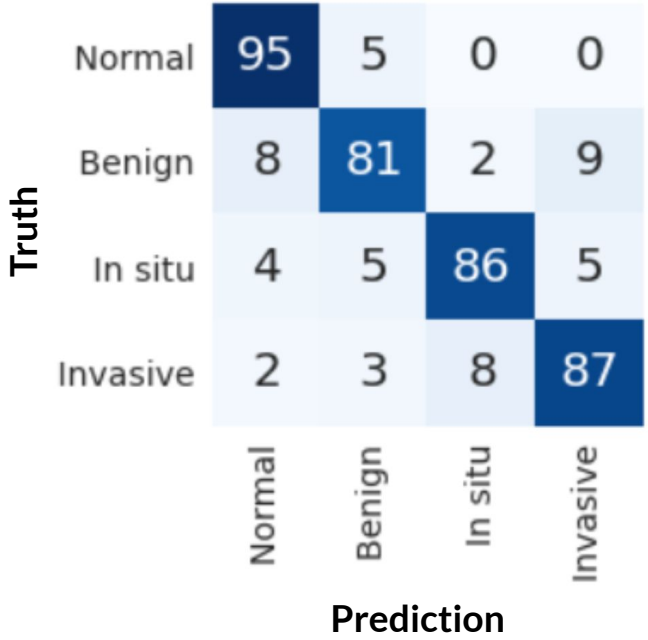
METHODOLOGY AND ANALYSIS (M&A)



M&A – MODEL TRAINING

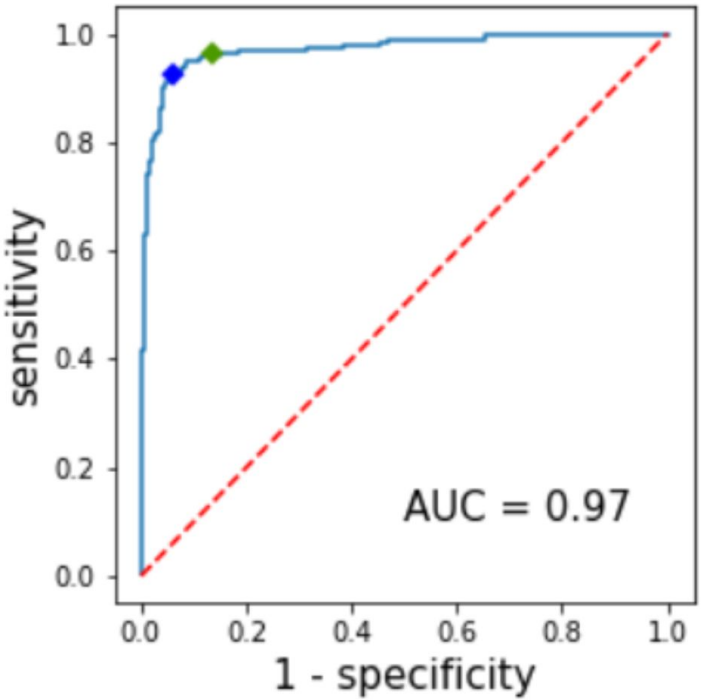
- Augmentations increase the size of the dataset x 300 (2 patch sizes x 3 encoders x 50 color augmentation)
- Cross validation with gradient boosting for classification -
  - 10 stratified folds to preserve class distribution
  - LightGBM - ensemble classifier
  - Each dataset recycled 5 times with different random seeds in LightGBM - augmentation at model level

M&A - Results



- For 4-class classification accuracy averaged across all folds was 87.2 ± 2.6%
- Out of 200 carcinomas cases only 9 in situ and 5 invasive were missed

M&A - RESULTS



- For 2-class non-carcinomas (normal and benign) VS carcinomas (in situa and invasive) classification accuracy 93.8 ± 2.3%
- The area under the ROC curve was 0.973
- ROC. 96.5% sensitivity at high sensitivity setpoint (green)

M&A – RESULTS

model	f 1	f 2	f 3	f 4	f 5	f 6	f 7	f 8	f 9	f 10	mean	std
ResNet-400	92.0	77.5	86.5	87.5	79.5	84.0	85.0	83.0	84.0	82.5	84.2	4.2
ResNet-650	91.0	77.5	86.0	89.5	81.0	74.0	85.5	83.0	84.5	82.5	83.5	5.2
VGG-400	87.5	83.0	81.5	84.0	84.0	82.5	80.5	82.0	87.5	83.0	83.6	2.9
VGG-650	89.5	85.5	78.5	85.0	81.0	78.0	81.5	85.5	89.0	80.5	83.4	4.4
Inception-400	93.0	86.0	71.5	92.0	85.0	84.5	82.5	79.0	79.5	76.5	83.0	6.5
Inception-650	91.0	84.5	73.5	90.0	84.0	81.0	82.0	84.5	78.0	77.0	82.5	5.5

LEARNINGS

- **Image Augmentations** - very important with sparse data
- **Light GBM hyperparameter tuning** -
  - **max_depth**
    - It describes the maximum depth of tree. This parameter is used to handle model overfitting. Any time you feel that your model is overfitted, my first advice will be to lower max_depth.
  - **feature_fraction**
    - Used when your boosting is random forest. 0.8 feature fraction means LightGBM will select 80% of parameters randomly in each iteration for building trees.

LEARNINGS

- **bagging_fraction**
  - specifies the fraction of data to be used for each iteration and is generally used to speed up the training and avoid overfitting
- **bagging_freq**
  - This parameter can help you speed up your analysis. Model will perform bagging at every k iterations mentioned. Provides good mix to data
- **num_leaves**
  - This is the main parameter to control the complexity of the tree model. **Ideally, the value of num_leaves should be less than or equal to 2^(max_depth)**. Value more than this will result in overfitting.

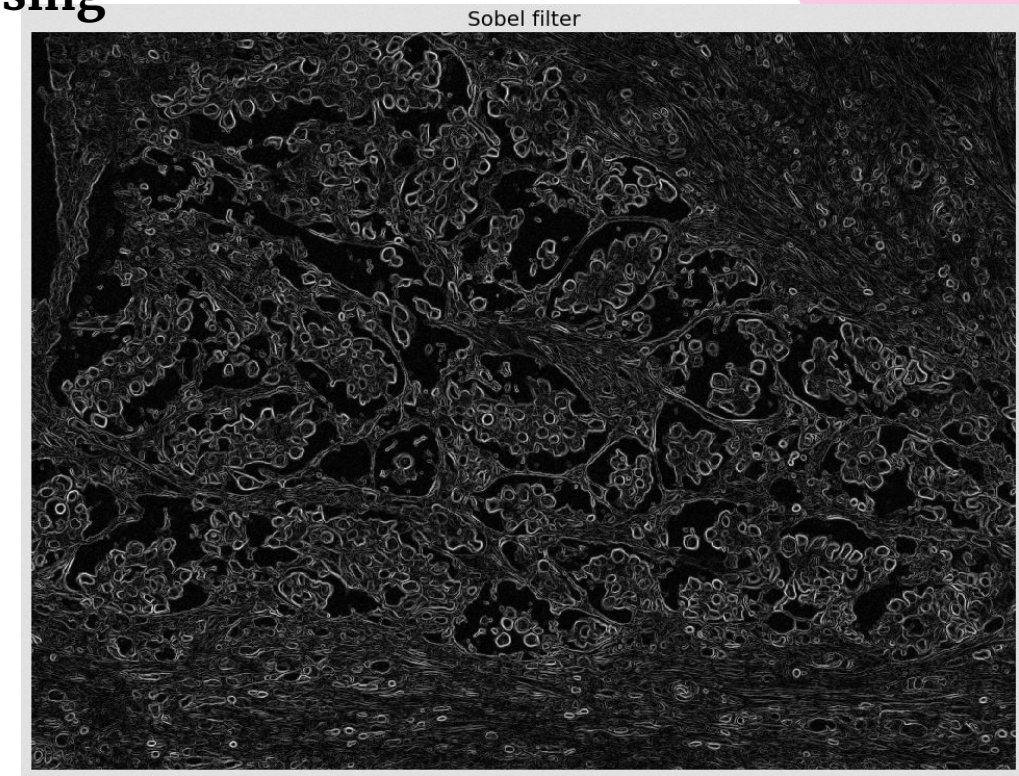
LEARNINGS

- **Neural Net Generalisation** - concept of dropout neurons (concept of ensemble)

Future Work: Processing Pipeline

Other forms of intelligent preprocessing

- ➔ **Edge Detection**
  - ◆ Markup Image?
  - ◆ Create meta-data:
    - Cell count/Density
    - Describe cell topology
      - homogeneity?
- ➔ **Localized metadata**
  - ◆ New convolution features?
  - ◆ Parallel convolutions?



Future Work: Analysis/Feature Processing

Select best network

Compare error types

- Accuracy vs Misclassification

Spin up Advanced Network

- Extract features/markup, merge w/ feature vector
- Implement Domain Knowledge?

Future Work: Interpretability

Create a model that can explain why it reached the decision it did

- ➔ Visualizations of decision-feature relevance
- ➔ Layer-wise Relevance Propagation

https://lrpserver.hhi.fraunhofer.de/image-classification

Thank You!

Any Questions?