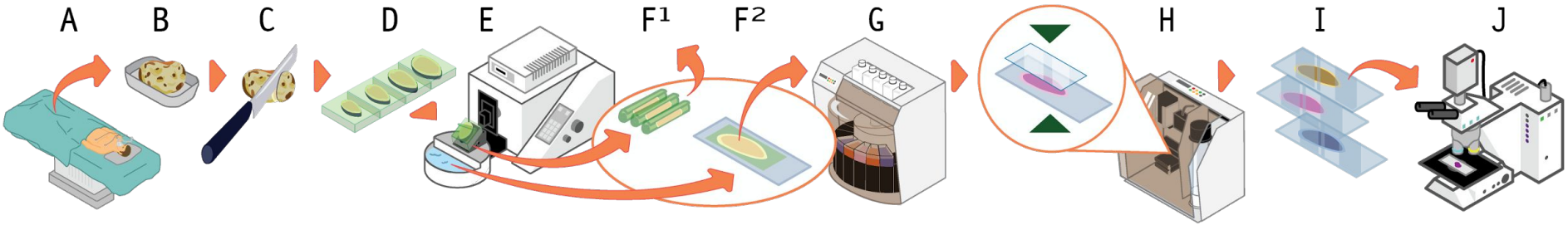
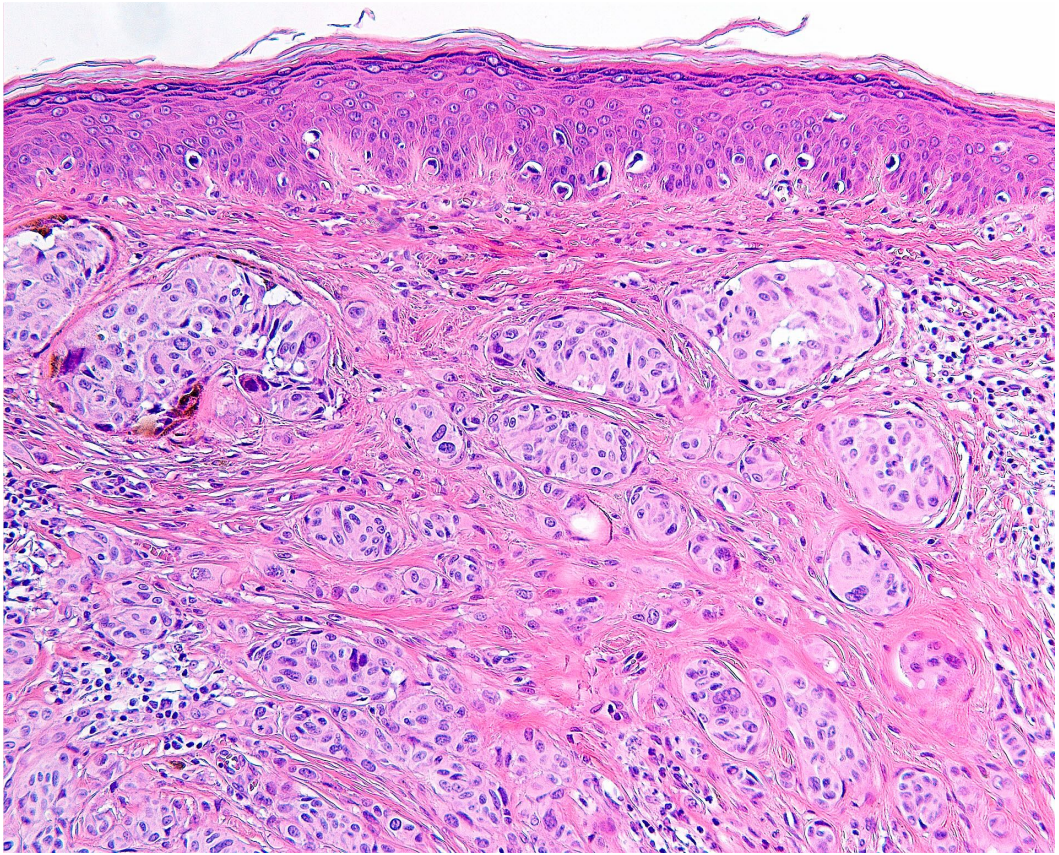
Deep Learning for Identifying Metastatic Breast Cancer

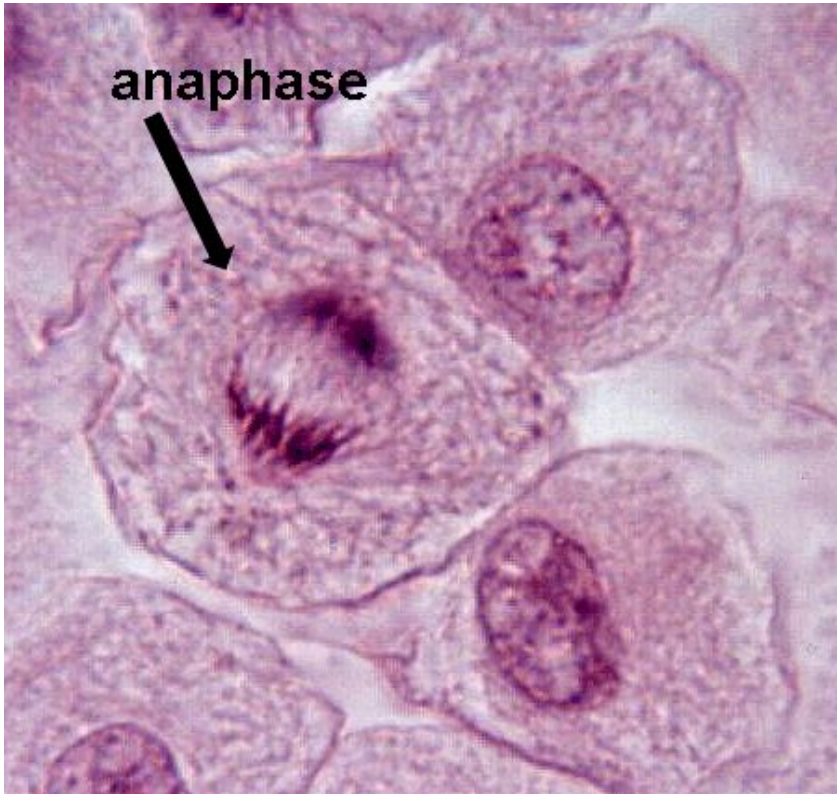
Naveen Arivazhagan, Jared Dunnmon, Carson Lam, Priyanka Nigam, and Darvin Yi

Introduction and Background

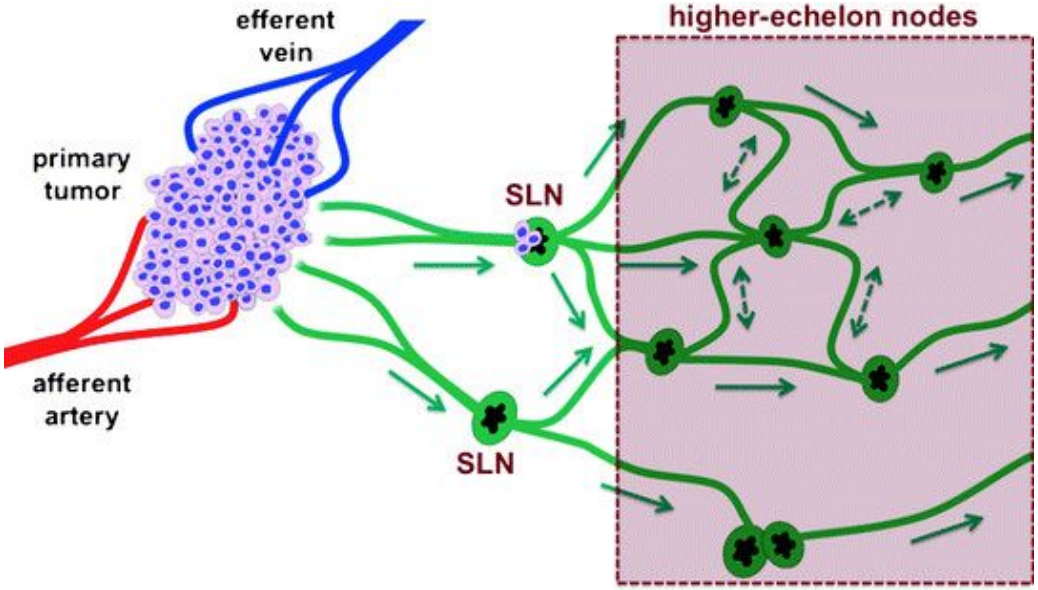
The Pathology Pipeline






anaphase

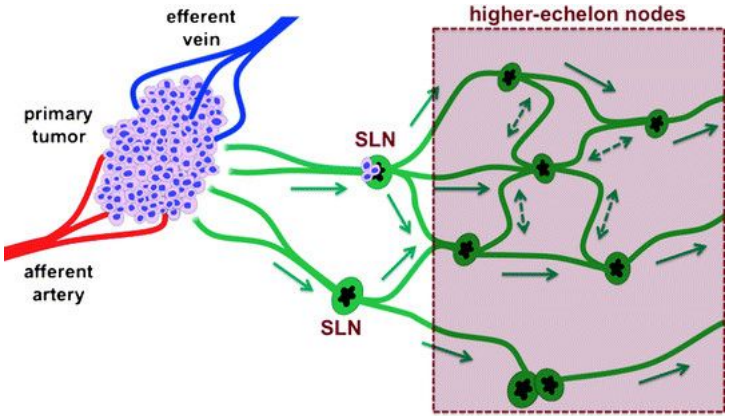
Sentinel Lymph Node



metastatic cancer goes here first usually

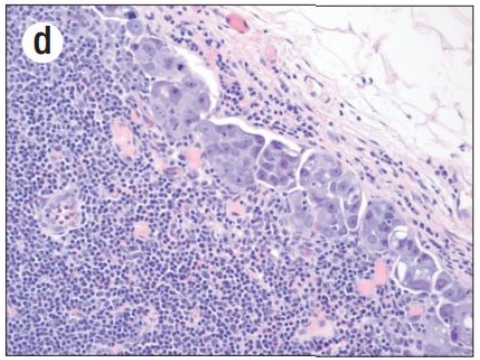
Sentinel Lymph Node Biopsies

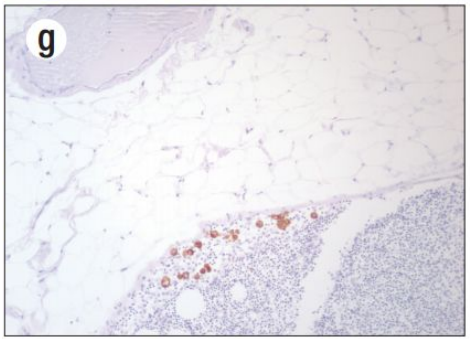
- Used in TNM Cancer Staging System
  - T - notes the size of the primary tumor
  - N - notes whether nearby lymph nodes are involved
  - M - notes presence of metastasis
  - Negative nodes suggests cancer has not yet spread to lymph nodes or organs nearby

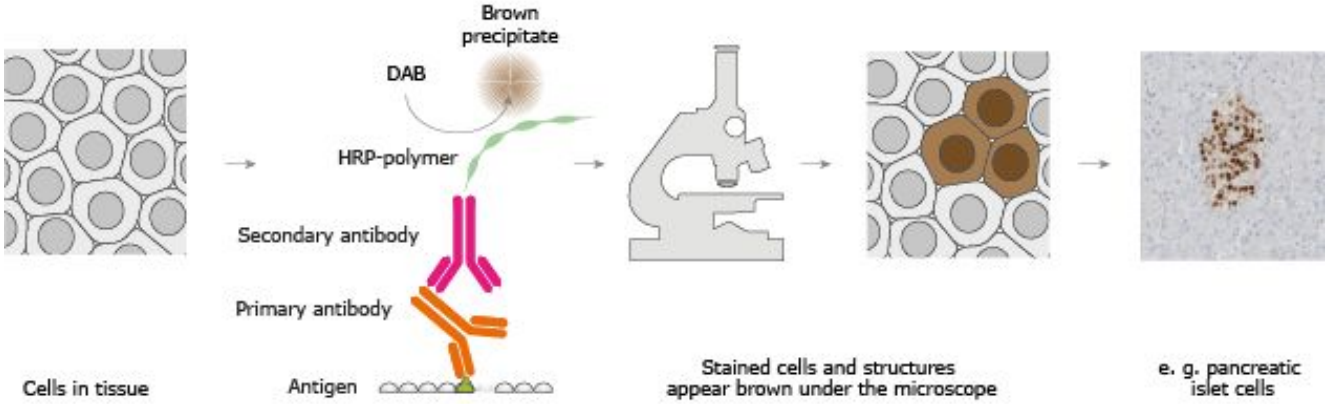


Immunohistochemistry to increase sensitivity

- Detecting presence of antigens in tissue by binding of antibodies
- Limitations
  - Increased cost
  - Increase sample preparation time
  - Increased number of samples to be manually review by pathologists
  - Can still be inaccurate with small metastases and false positives







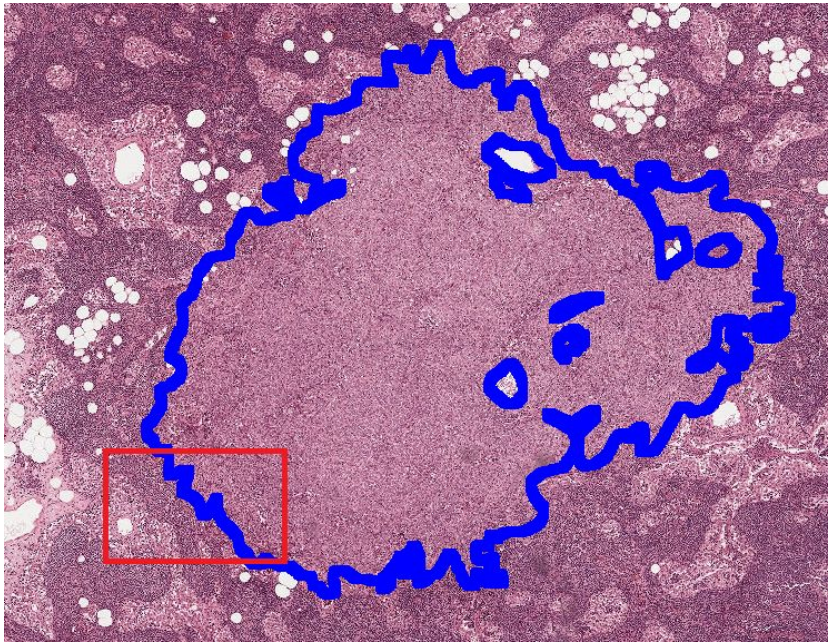
Key Limitations

- Lack of standardization (large inter-reader variability ~10 - 15%)
- Current augmentation: Immunohistochemistry with cytokeratin AE1-3
  - 88% sensitivity for macrometastasis and a 72% sensitivity for micrometastasis
- Time required to manually prepare and evaluate samples (days to weeks!)

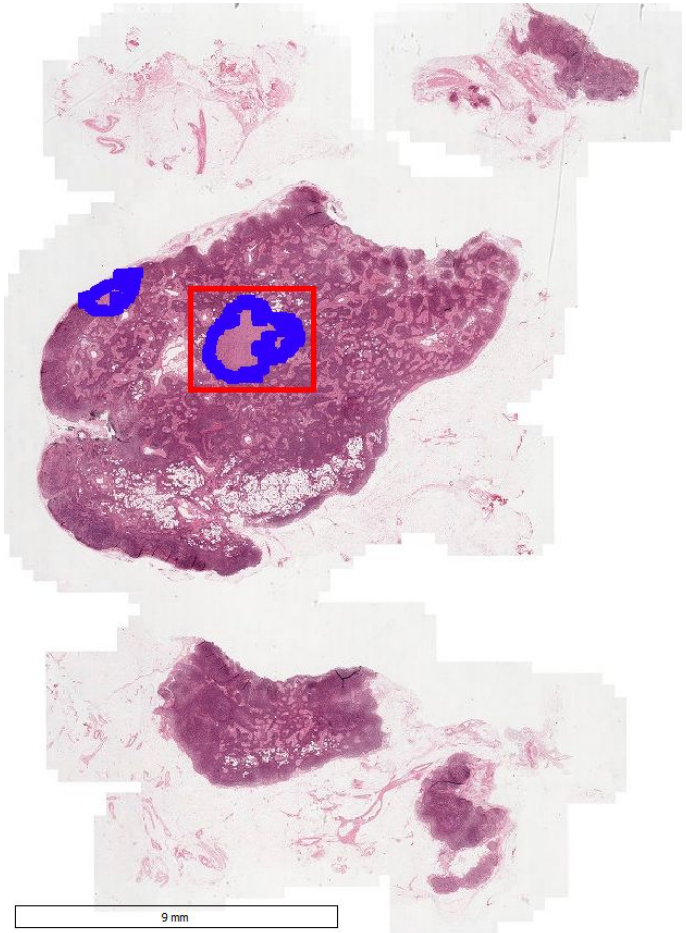
Meier, Frederick A. "The Landscape of Error in Surgical Pathology." Error Reduction and Prevention in Surgical Pathology. Springer New York, 2015. 3-26.

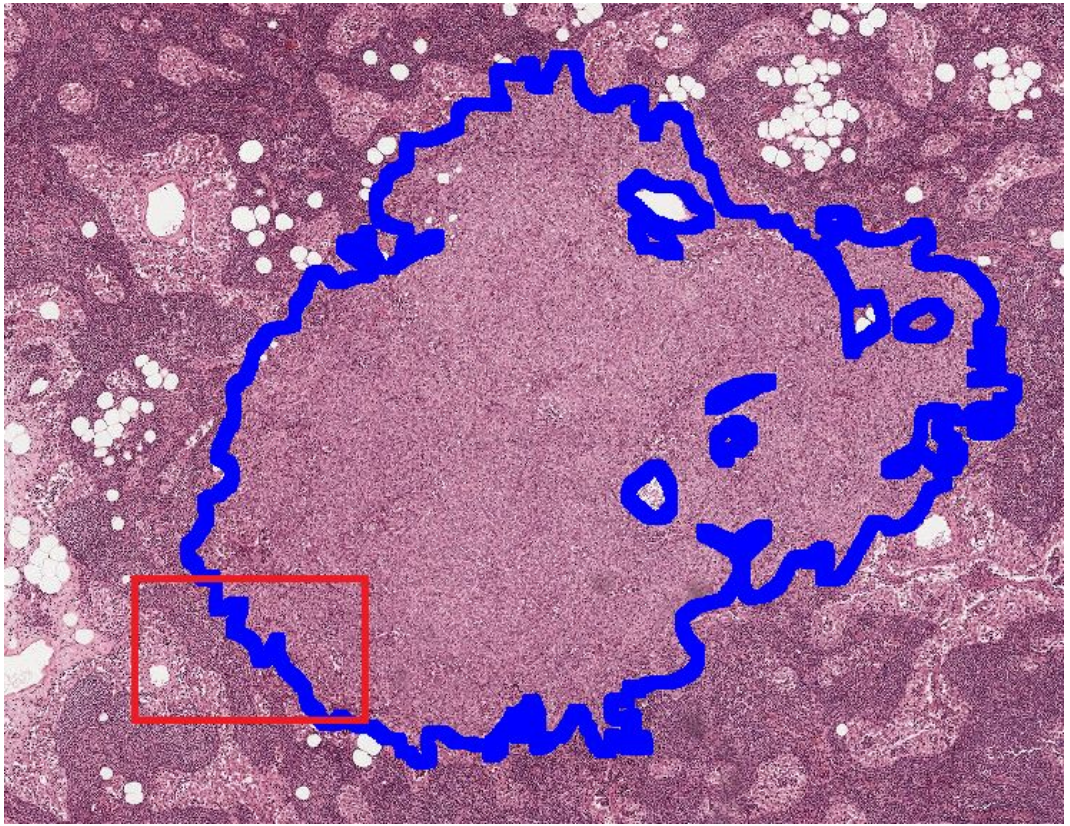
Camelyon Grand Challenge

- Task: Automate identification of metastatic breast cancer in hematoxylin and eosin (H&E) stained whole slide images of sentinel lymph node biopsies

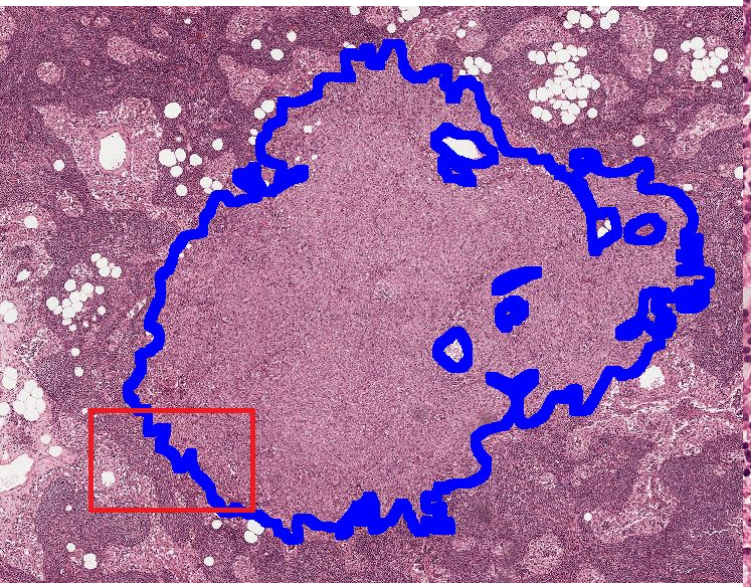


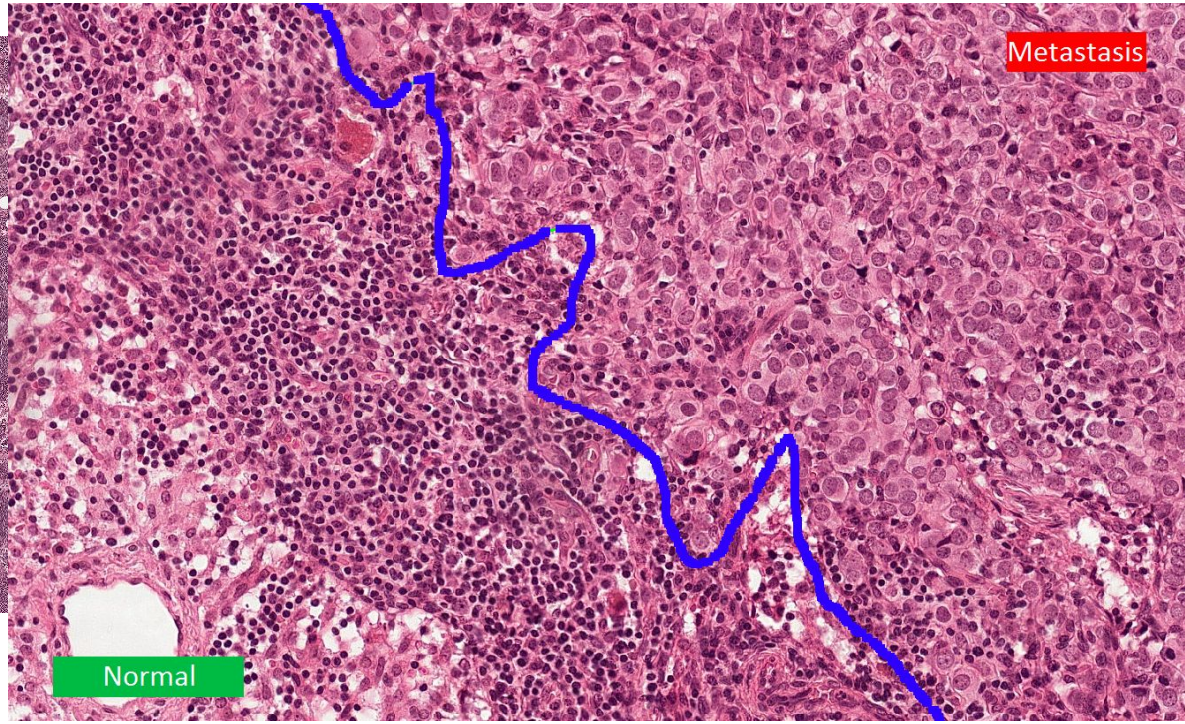
Camelyon Grand Challenge - 1x & 10x





Camelyon Grand Challenge 10x & 40x





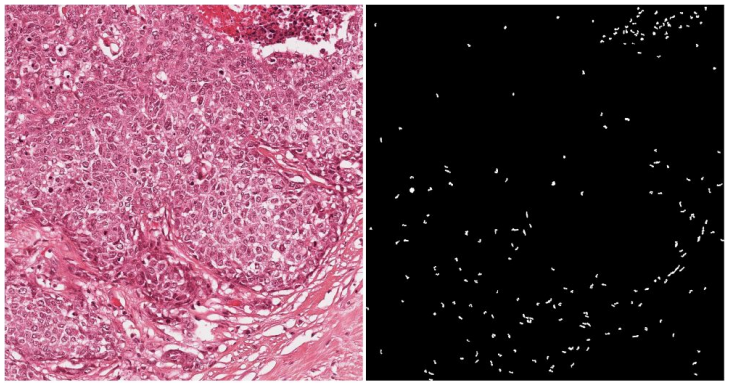
Camelyon16 Dataset - 400 high res images of the whole slide - google maps level of data per slides

- Data from 2 sites: Radboud University Medical Center and University Medical Center Utrecht (Netherlands)
- 2 training sets (270 total)
- 100 normal and 70 slides containing metastasis in 1st set
- 60 normal and 40 slides containing metastasis in 2nd set
- 130 images in test set

Camelyon16 Dataset Labels

The ground truth data for the slides containing metastases is provided in two formats:

- .xml files containing vertices of the annotated contours (corners of the blue borders)
- WSI metastasis binary masks

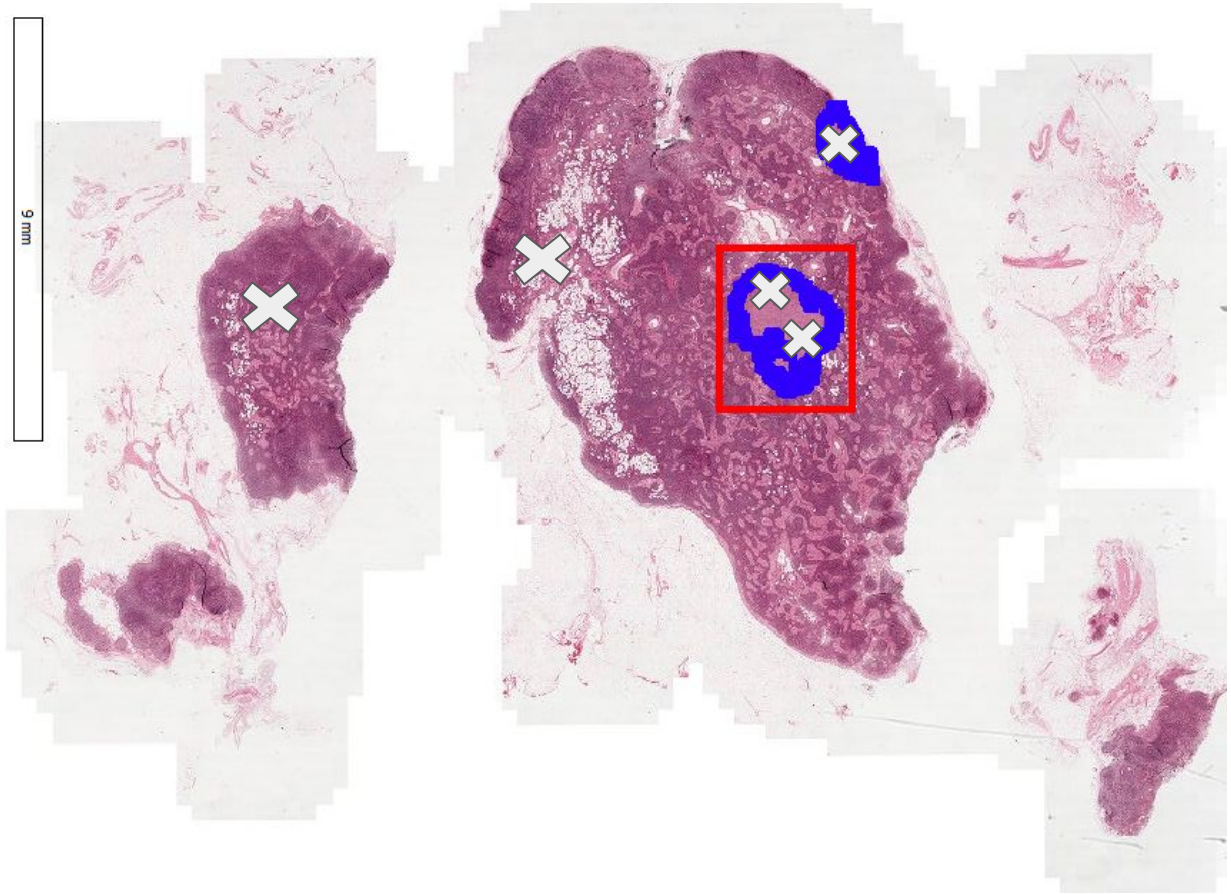


(a) (b)

Slide-Based Classification

image name	probability of disease
Slide 1	0.34
Slide 2	0.89
Slide 2	0.45
Slide 3	0.57

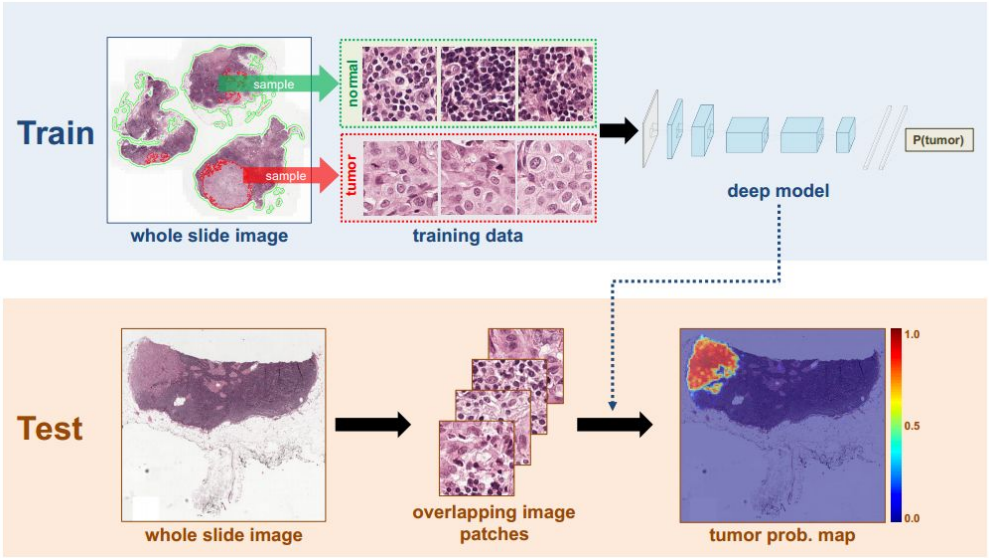
Lesion (~Segmentation) based evaluation



Technical Approach Overview

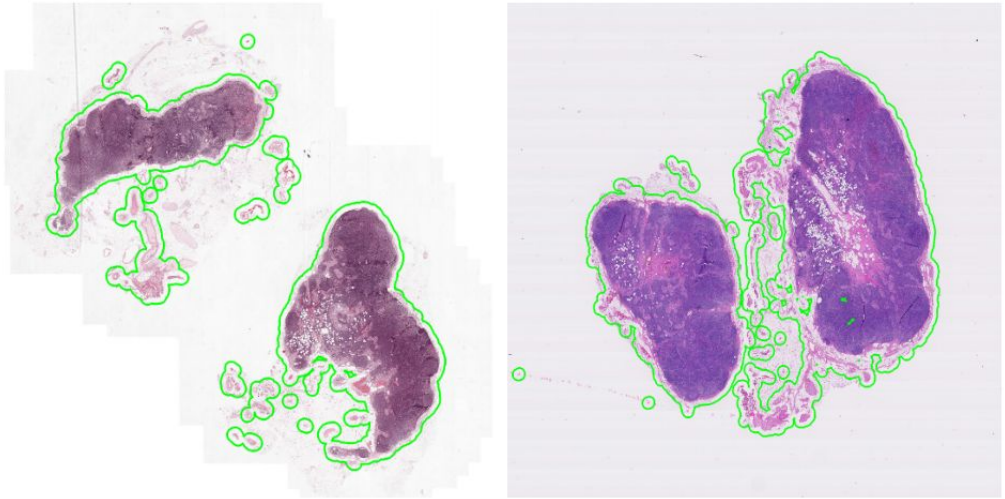
Data Processing Framework

- Step 1: Tissue and background identification
- Step 2: Patch-based classification
- Step 3: Heatmap-based post processing



Tissue Detection

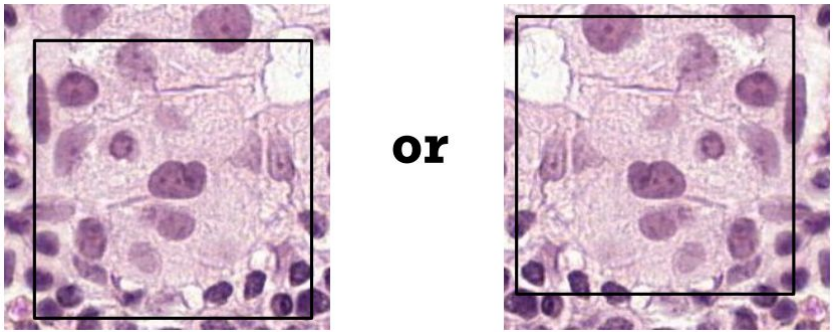
- RGB to HSV (hue, saturation, value) color space
- Otsu algorithm to determine threshold values for each channel
- Combine mask from H and S channels



Patch-based Classification

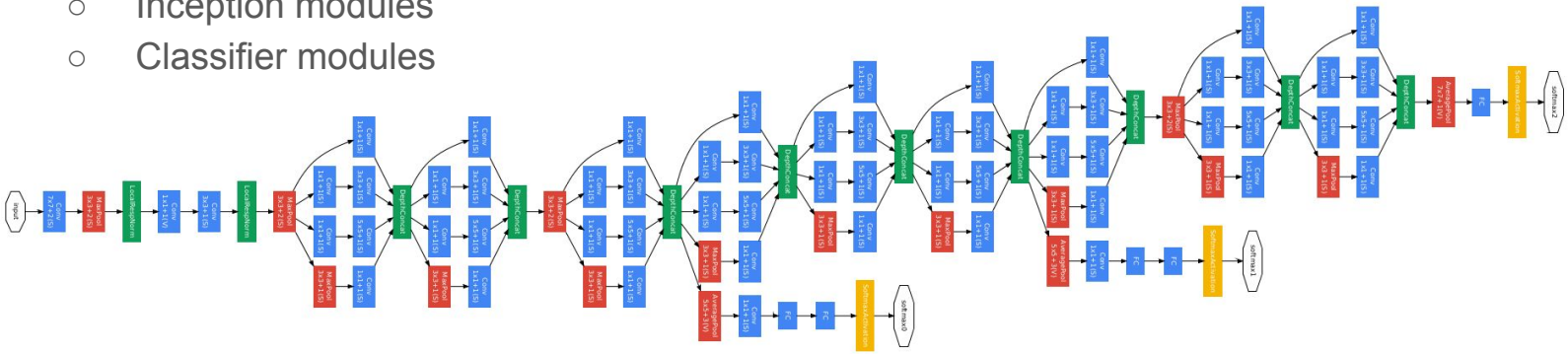
- Input: 256 x 256 pixel patches
- Evaluated several well-known architectures: **GoogLeNet**, AlexNet, VGG16, FaceNet
- Magnification levels: 10X, 20X, **40X**
- Data Augmentation - randomly crop a 224x224 region and flip horizontally

	Patch classification accuracy
GoogLeNet [20]	98.4%
AlexNet [12]	92.1%
VGG16 [19]	97.9%
FaceNet [21]	96.8%



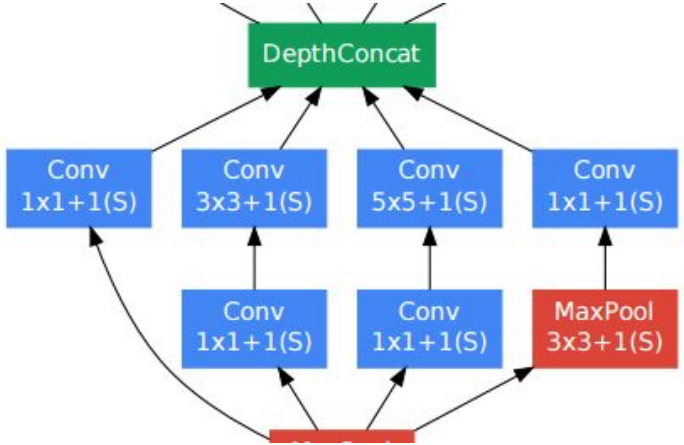
GoogLeNet Architecture

- 27 Layers
- 6 million parameters
- Broadly successful in image recognition tasks
- Several different layer architectures
  - Stem
  - Inception modules
  - Classifier modules



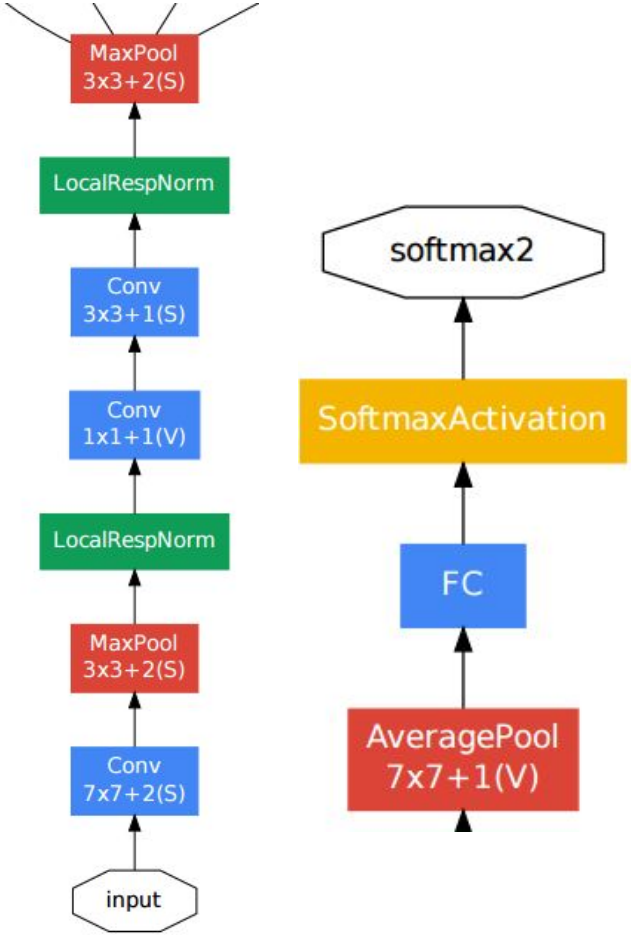
Inception Modules

- 9 inception modules in network
- Filters of different sizes
- Cover large area but keep finer resolution
- High performing model with fewer parameters
- Multiple different kernel sizes enable simultaneous analysis of features at several relevant scales



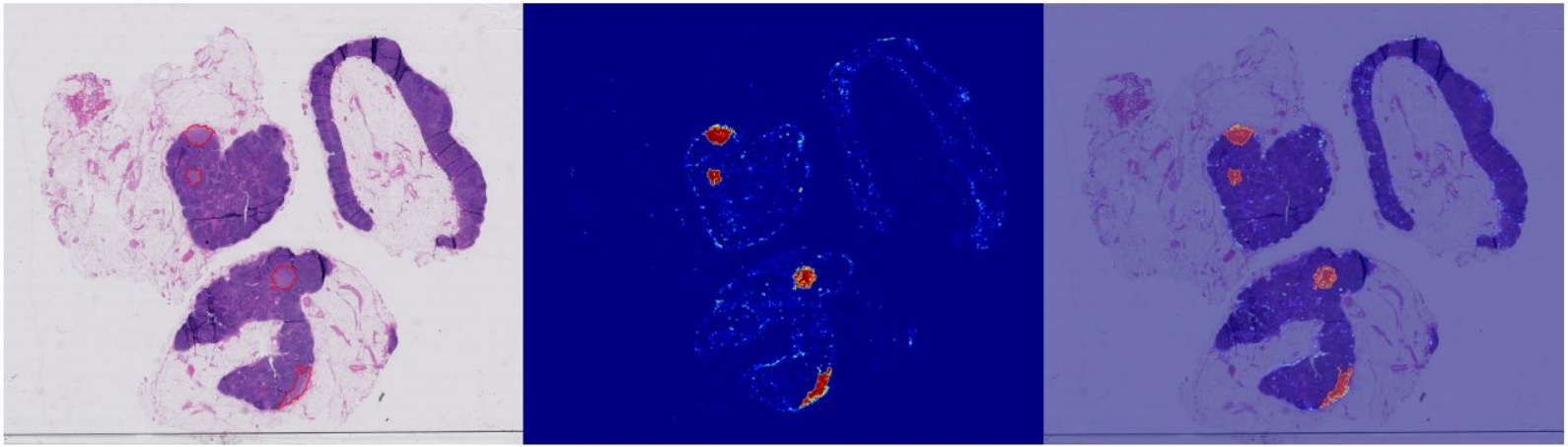
Stem and Output Classifier

- Stem
  - Several initial convolutional and pooling layers establish features that are input into multiple inception layers
  - Local response normalization -- scales output by taking nearby pixels into account
- Classifier
  - Wide-field average pooling before softmax and fully connected layer
  - Each inception layer can have a different classification output



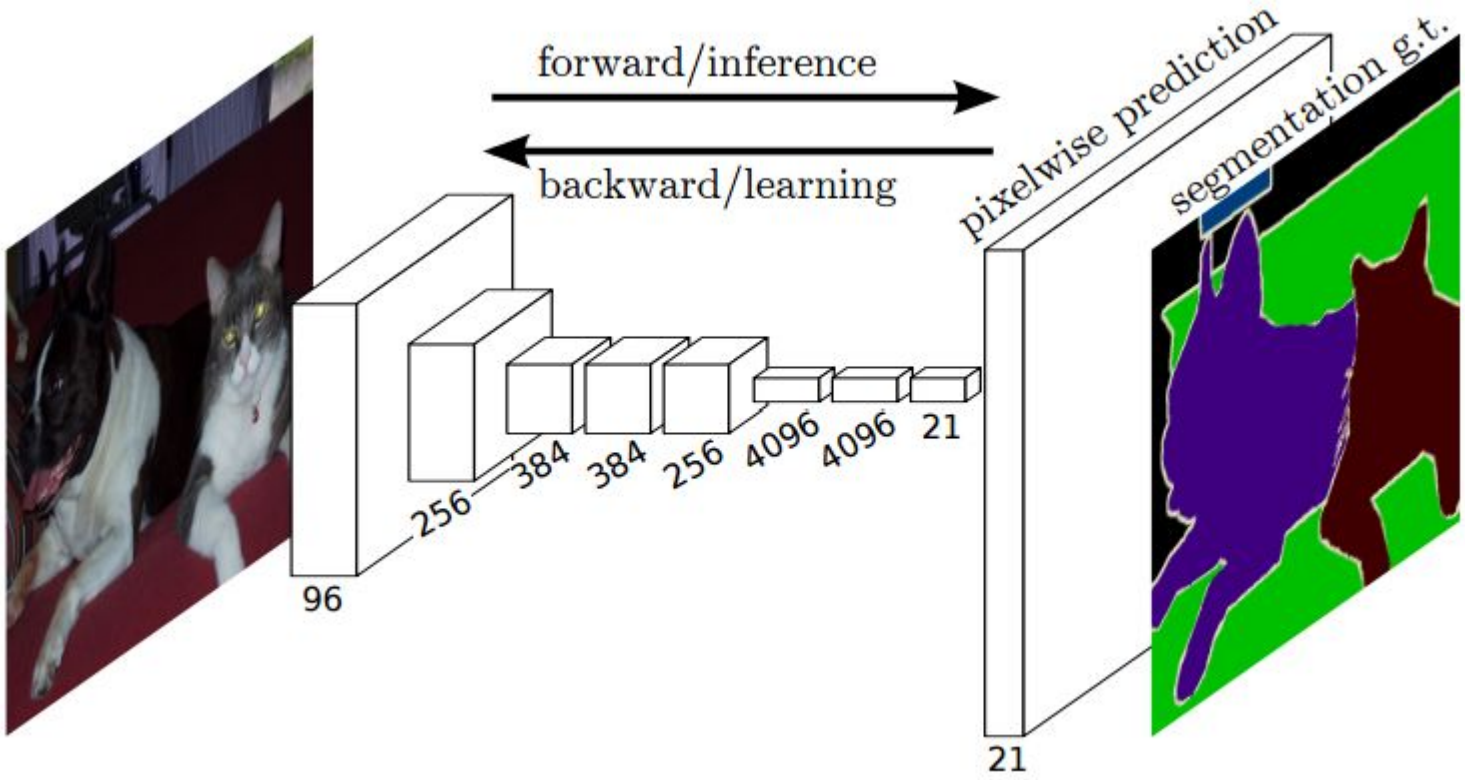
Heatmap Generation

- Heatmap - each pixel indicates the probability that the pixel contains a tumor
- Embed results for each patch into single heatmap for the entire slide



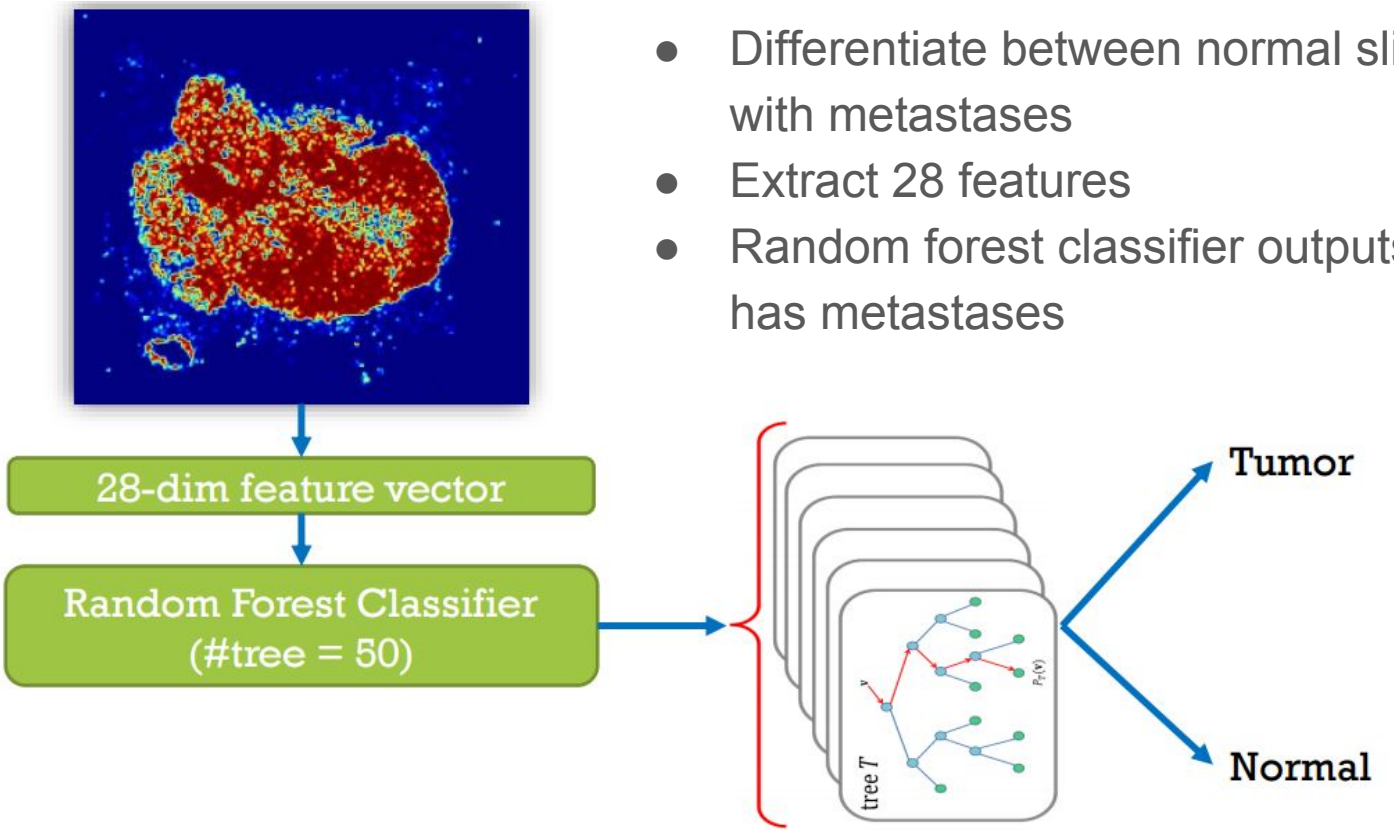
(a) Tumor Slide (b) Heatmap (c) Heatmap overlaid on slide

Patch-wise “Segmentation” vs. Fully Convolution Net



Classification Methods and Results

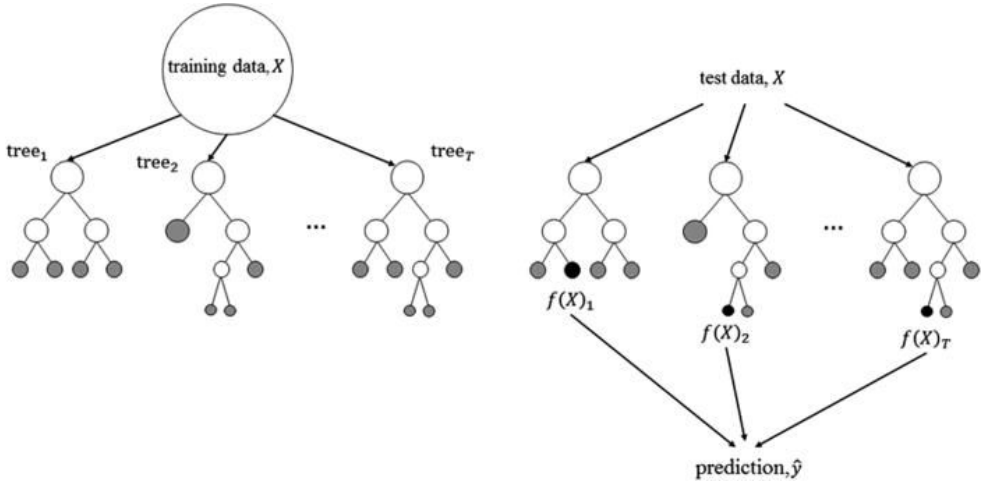
Slide-Based Classification



- Differentiate between normal slides and slides with metastases
- Extract 28 features
- Random forest classifier outputs probability WSI has metastases

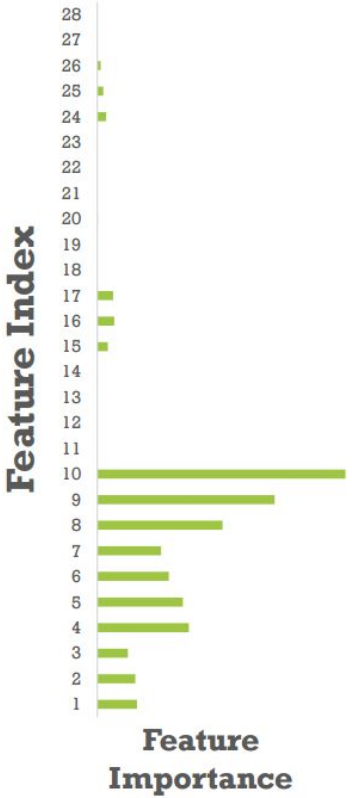
Random Forest Classifier

- Decision Tree
  - Each node represents a test on a feature
  - Outgoing branches then indicate the path to be taken based on the feature test
- Random forest ensemble method
  - Multiple trees over random subset of data to reduce overfitting
  - Parameters: Number of decision trees and depth of decision trees
  - Prediction: Most common classification by all trees

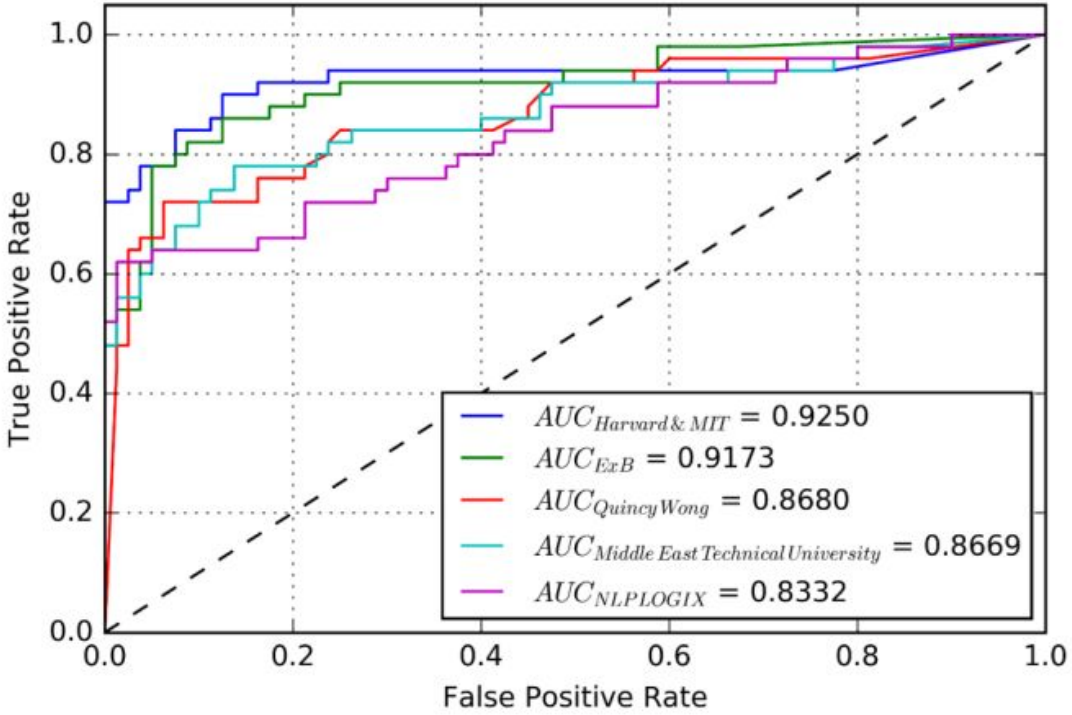


Most Important Features

- Feature 10: the longest axis in the largest tumor region
- Feature 09: ratio of pixels in the region to pixels in the total bounding box
- Feature 08: eccentricity of the ellipse that has the same second–moments as the region
- Feature 04: ratio of tumor region to the tissue region
- Feature 05: the area of largest tumor region



Slide-Based Classification Results



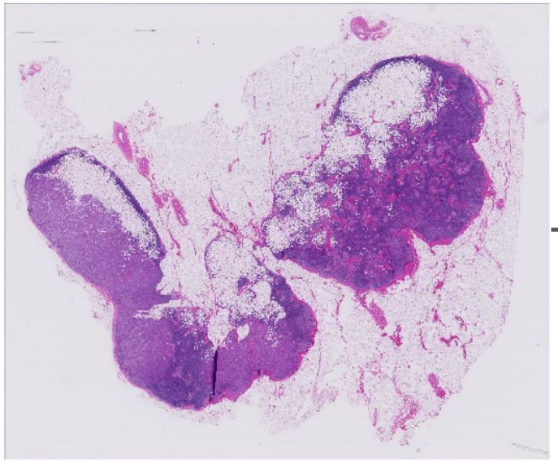
- Metric: Area Under the Receiver Operating Curve (AUC)
- Better prediction at low false positive rates
- Pathologist AUC was 0.9664

Other Top Performing Methods

- ExB
  - Tile-wise binary classification using 34-layer deep residual networks
  - Experimented with depths of 18 to 101 layers
- Quincy Wong
  - Used a pretrained VGG-16 model, 37 layer, fully convolutional
- Middle East Technical University
  - 64 x 64 randomly selected patches with < 75% background
  - Network - 2 convolutional layers for feature extraction, 2 fully connected layers and a softmax for classification
- NLPLOGIX
  - 7-layer neural network - first 5 layers are convolutional, the last 2 layers are fully connected
  - 256 x 256 image patches randomly rotated and flipped

Lesion-Based Detection

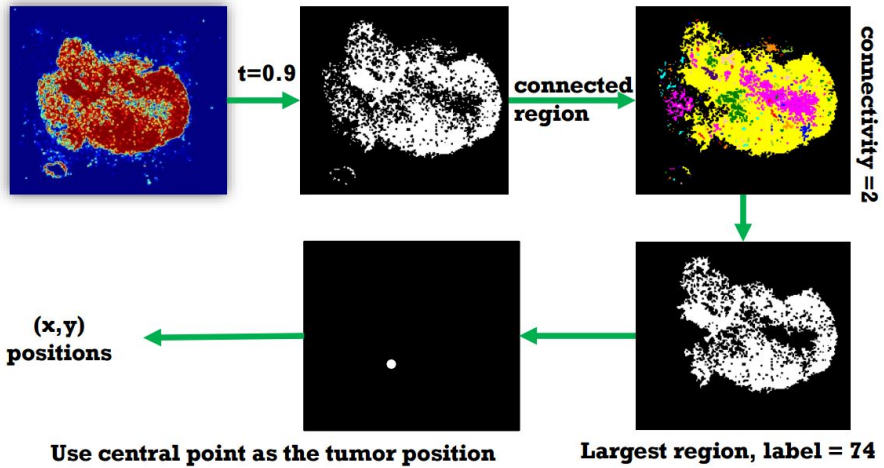
- Predict x and y coordinates of each tumor region and a confidence score for each region
- D-I – model trained on original data has greater sensitivity
- D-II – model trained on enriched data has greater specificity
  - 30,000 additional patches for tumor-adjacent false positive regions

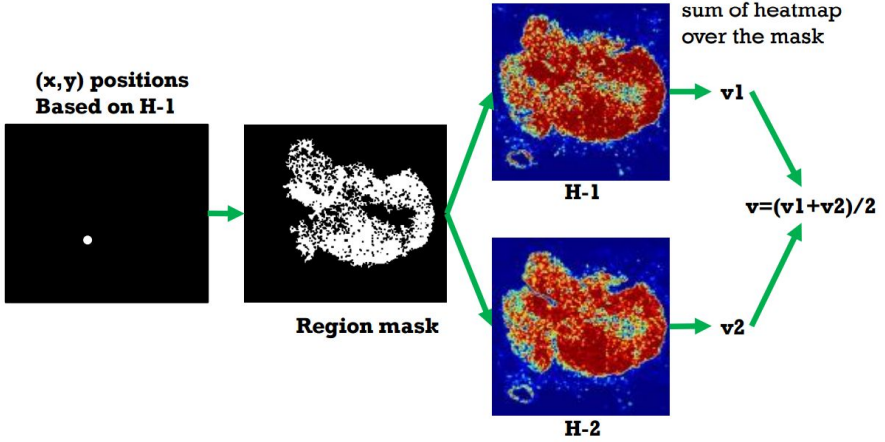


Confidence	X coordinate	Y coordinate
0.73	18298	169828
0.84	10498	165754
0.57	12122	153638
0.91	10866	154596
0.32	13742	121722
0.21	12458	134585
0.64	14250	146531

Lesion-Based Detection

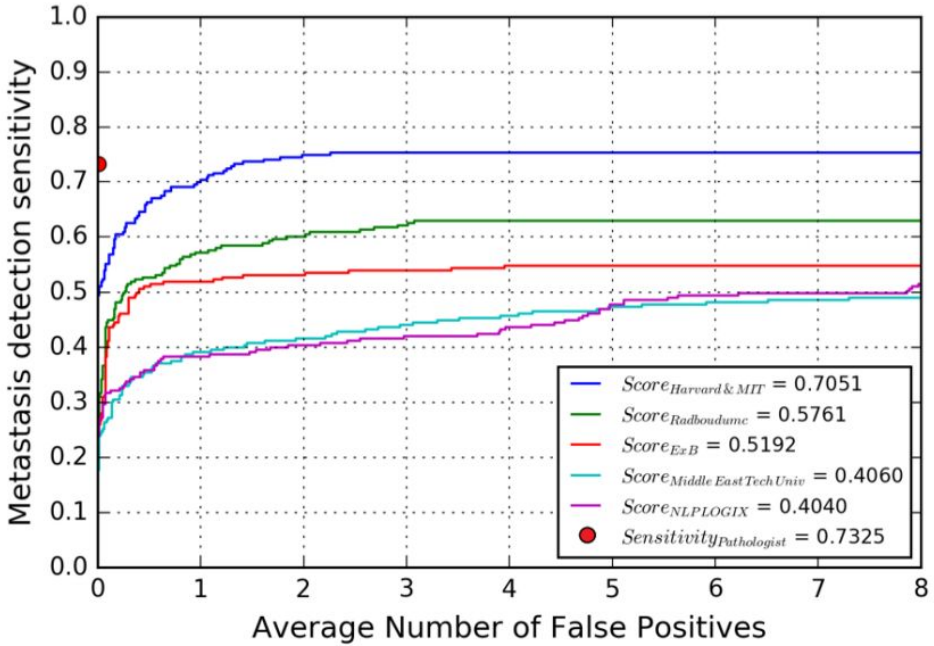
- Tumor localization
  - Use the heatmap generated by D-I to determine tumor location
- Confidence scores
  - Average heatmaps generated by D-I and D-II for prediction probabilities





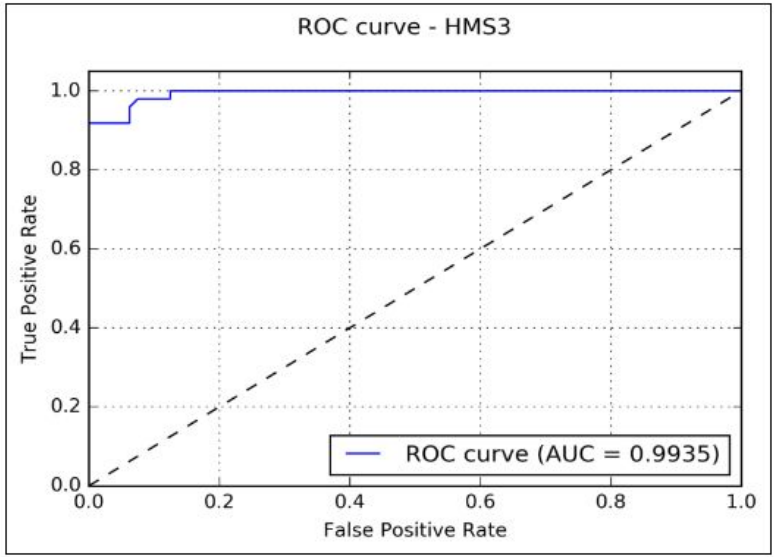
Lesion-Based Detection Results

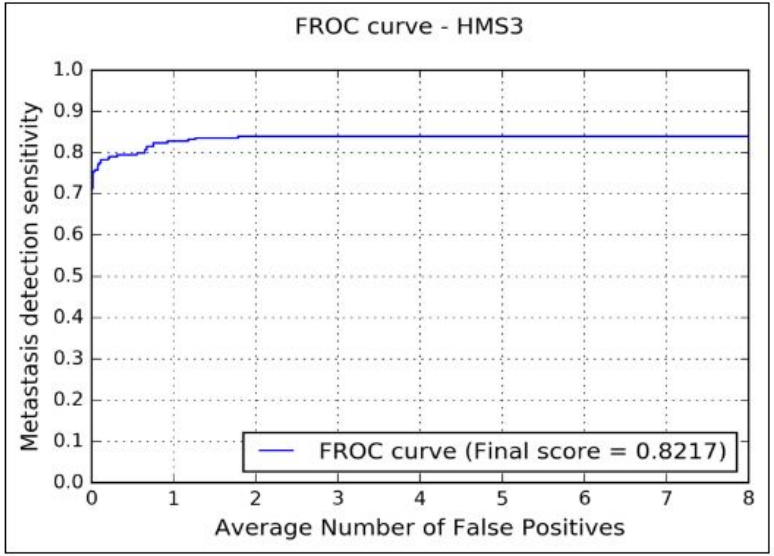
- Free Response Receiver Operating Characteristic (FROC)
- Scoring Metric: Average sensitivity at false positive rates of ¼, ½, 1, 2, 4, and 8
- Outperformed pathologist when the average number of false positives is greater than 2



Updated Method by Same Authors

- Same framework, different preprocessing methods
- Standardize all the images to have the same staining color
- Data augmentation: randomly rotated training patches with 90, 180, or 270 degrees and added extra color noise





Critiques and Conclusions

Selected Critiques and Followup Questions

- **Generalization:** how well does this pipeline work outside of metastatic breast cancer?
- **Ground truth:** ground truth is “a” pathologist, same with test comparison
- **Negative example selection:** are histologic mimics truly benign, or are they too difficult for a pathologist to differentiate?
- **Negative example integration**: averaging network results for DI and DII networks is a coarse approach to negative example mining
- **Workflow integration:** what regulatory and workflow hurdles need to be overcome to integrate this type of Computer Aided Detection into physician workflows?
- **Knowledge extraction**: can network visualization demonstrate learning of clinically relevant topological features?

Selected Critiques and Followup Questions

- **Direct DL classifier:** While Random Forests seem to perform well on the WSI classification task, could a direct DL-based classifier be used?
- **Utilizing targeted preprocessing:** augmenting features from clinically-relevant preprocessing could improve results -- future work should investigate the cost-benefit analysis of adding specific preprocessing steps
- **Naive thresholding:** thresholding approaches such as the 0.9 probability threshold on heat map binarization require human input -- automated methods would be superior, and potentially more robust in practice
- **Quantifying impact:** how well could this CAD system decrease misread rate in a clinical setting? What is the effect on the best human versus the average?
- **Runtime reporting:** how long did this take to train and run?

Summary of Main Contributions

- **Achieve near-human performance on WSI sentinel lymph node pathology with a deep CNN architecture**
  - Multiple architectures give accuracies of over 95% on patch classification
  - WSI classification score of 0.925 AUC, localization score of 0.733
  - No requirements for color normalization, nuclear segmentation, or explicit feature design in classification -- major relief of burden on human analysts and designers
- **Suggest efficacy of negative example mining in improving performance**
  - Augmenting datasets with non-tumor histologic mimics of cancer improves classification performance
- **Demonstrate independence of pathologist and DL errors**
  - Pathologist AUC: 0.9664, DL AUC: 0.925 -- note that this is the *best* a pathologist could do, not the average pathologist performance in a clinical environment
  - **Combined human-DL AUC 0.9948 -- this is very encouraging for how DL can actually help pathologists!**

Moving Forward...Commercialization?

https://www.pathai.com/

Questions?

References

1. The Human Protein Atlas. “Immunohistochemistry.” http://www.proteinatlas.org/learn/method/immunohistochemistry.
2. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. Going deeper with convolutions. *CoRR*, abs/1409.4842, 2014.
3. Shiller SM, Weir R, Pippen J, Punar M, Savino D. The sensitivity and specificity of sentinel lymph node biopsy for breast cancer at Baylor University Medical Center at Dallas: a retrospective review of 488 cases. *Bayl Univ Med Cent Proc.* 2011;24:81–85.
4. Meier, Frederick A. "The Landscape of Error in Surgical Pathology." Error Reduction and Prevention in Surgical Pathology. Springer New York, 2015. 3-26.
5. https://camelyon16.grand-challenge.org/site/CAMELYON16/serve/public_html/Results/Presentations/Camelyon16_BIDMC_CSAIL.pdf/
6. Guidance for Instrudy and Food and Drug Administration Staff - Computer Assisted Detection Devices Applied to Radiology Images and Radiology Device Data - Premarket Notification [510(k)] Submissions. United States Food and Drug Administration. Issued July 3, 2012.