

# Cervical Cytology Image Analysis and Classification Using Graph-Based Techniques and Deep Learning

Sudhir Sornapudi, Ph.D. student<sup>1</sup>, Gregory T. Brown, M.D., Ph.D.<sup>2</sup>, Zhiyun Xue, Ph.D.<sup>2</sup>, Rodney Long, M.A.<sup>2</sup>, Lisa Allen, B.S., CT<sup>3</sup>, Sameer Antani, Ph.D.<sup>2</sup>

<sup>1</sup>Missouri University of Science and Technology, Rolla, MO, USA; <sup>2</sup>Lister Hill National Center for Biomedical Communications, U.S. National Library of Medicine, Bethesda, MD, USA; <sup>3</sup>Diagnostic Systems Women's Health and Cancer, Becton Dickinson and Company, Durham, NC, USA

**Sudhir Sornapudi**

Website: <https://sahir.github.io>

Twitter: @Sahir23



**BD**

S54: Oral Presentations

Translational Bioinformatics and Big Data in Cancer

#AMIA2019



# Disclosure

---

The authors have no commercial interests to disclose.

Those with BD are already commercial, but the work wasn't commercial in nature. Included all authors and affiliations on main slide.

# Learning Objectives

---

After participating in this session the learner should be better able to:

- Understand the importance of image preprocessing automate the data preparation.
- Understand the concepts of image registration, superpixel, region adjacency graph, image classification.

## ➤ Cervical Cancer

- One of the most common cancer among women.
- 2018 Worldwide Statistics <sup>Δ</sup>
  - 570,000 new cases.
  - 311,000 women died from cervical cancer.
- Cervical cancer that is detected early is more likely to be treated successfully.
- Early detection is accomplished via cervical cytopathology from Pap smears.

<sup>Δ</sup> "Human papillomavirus (HPV) and cervical cancer," World Health Organization 2019.

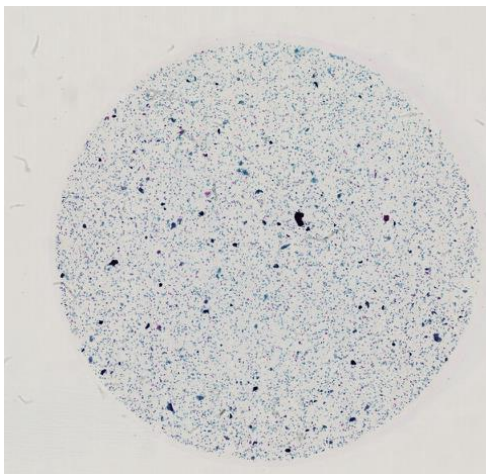
## ➤ NLM Data

- 25 cytology whole slide images.
- Provided by BD (Becton-Dickinson) Corporation.
- The slides are prepared through Liquid based Cytology (LBC).
  - Thin layer slide preparation technology
  - Using Sure Path technique

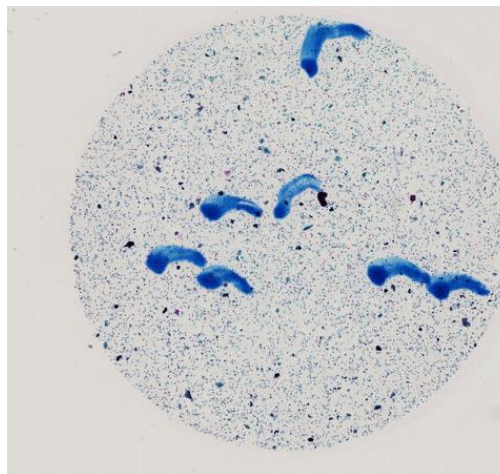
## ➤ Herlev Pap Smear Dataset

- 917 cervical cell images.
- Extracted through Conventional Pap Smear.

Clean Slide image



Annotated Slide image



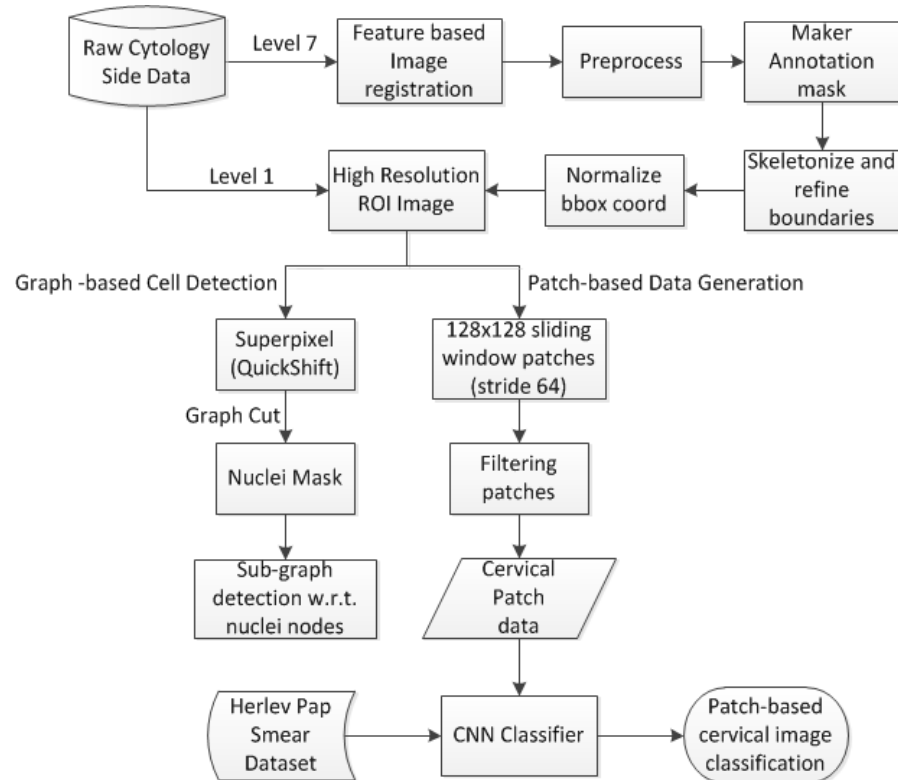
Note: Displayed images are from level 7

Image Properties

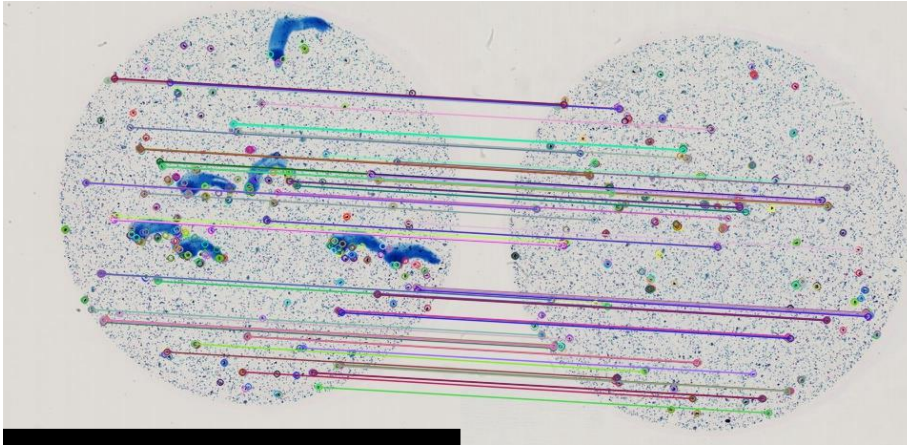
File Properties Macro image

Image type	Hamamatsu NDPI
File name	12XS00692.ndpi
File path	[REDACTED]
File size	342.2 MiB
Image width	61440
Image height	59136
Pixel size	0.228003 x 0.228003 um
Magnification	40X
Display color	RGB (32 bpp)
Image Centre	5.95953, 0.704433 mm
Pyramid structure	Level 0: (61440,59136) Level 1: (30720,29568) Level 2: (15360,14784) Level 3: (7680,7392) Level 4: (3840,3696) Level 5: (1920,1848) Level 6: (960,924) Level 7: (480,462) Level 8: (240,231)

# Framework of Proposed Approach



# Region of Interest (ROI) detection



## Matching Key-points

### ➤ Image Registration

- ORB feature detector <sup>Δ</sup>
- Match features
- Calculate Homography
  - Uses RANSAC\*  
estimation technique

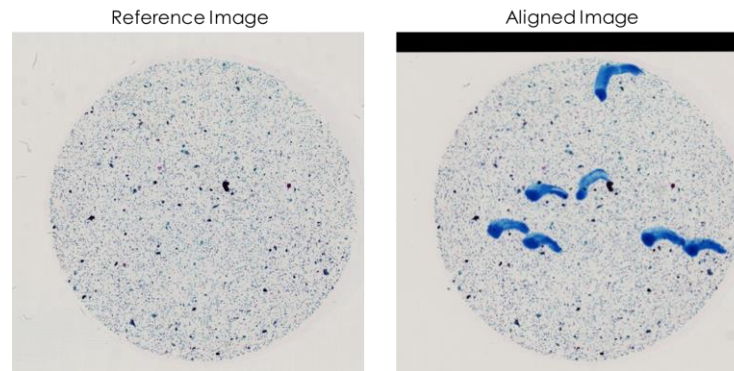
<sup>Δ</sup> Ethan Rublee, Vincent Rabaud, Kurt Konolige, Gary R. Bradski: ORB: An efficient alternative to SIFT or SURF. ICCV 2011: 2564-2571

\* Random sample consensus (**RANSAC**) is an iterative method to estimate parameters of a mathematical model from a set of observed data that contains outliers

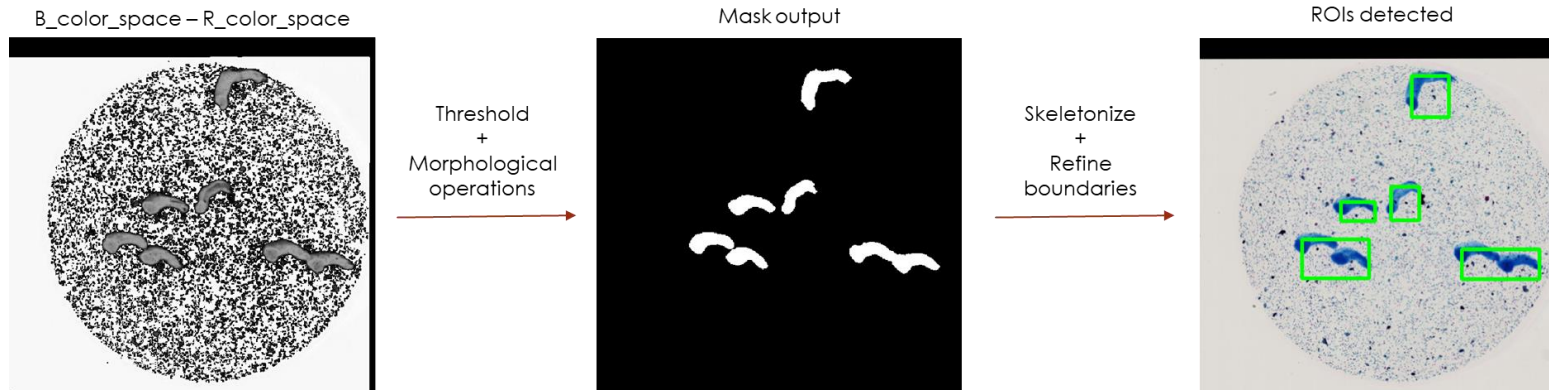


# Region of Interest (ROI) detection

## Image Registration



## Detecting marked regions



# Locating Abnormal Cells

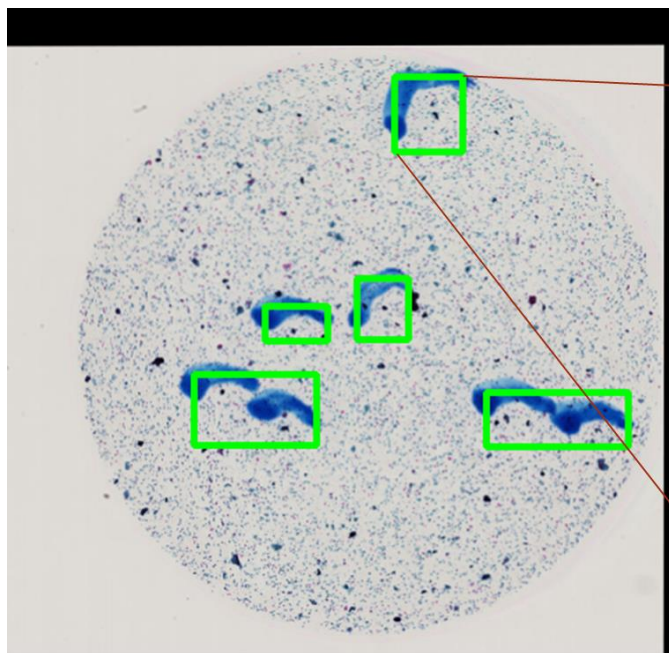
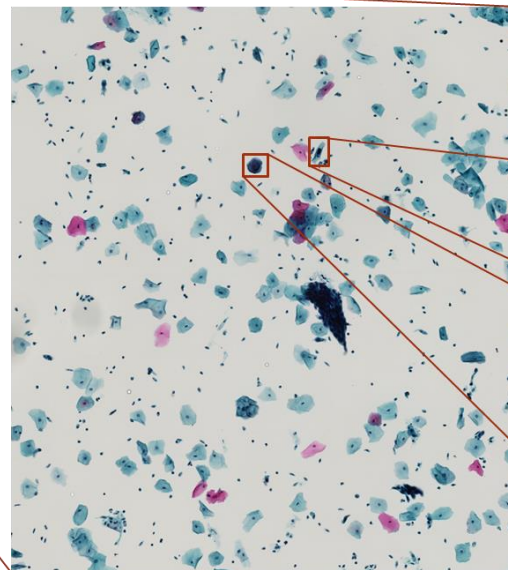


Image from level 7

Size: 3400x3079

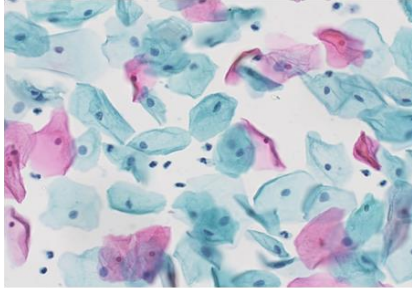


Abnormal cells

Image from level 1

# Cell Classification

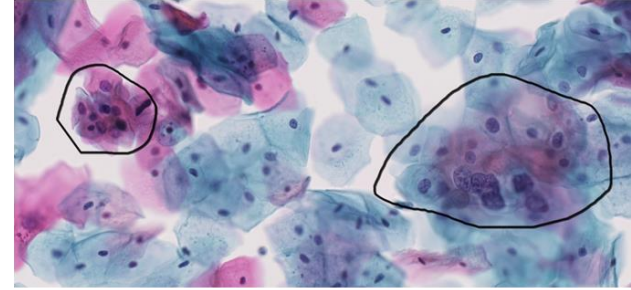
- NILM (Negative for Intraepithelial Lesion or Malignancy)



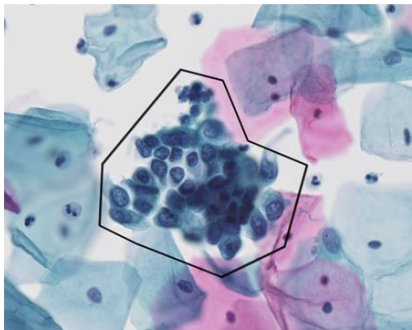
- ASCUS (Atypical squamous cells of undetermined significance)



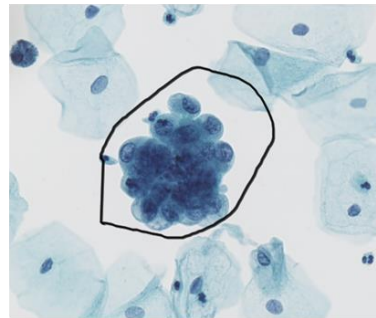
- LSIL (Lower-grade Squamous intraepithelial lesion)



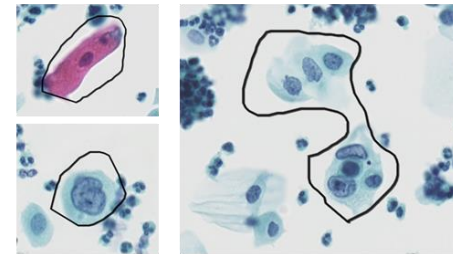
- HSIL (Higher-grade Squamous intraepithelial lesion)



- Adeno (Adenocarcinoma)



- SCC (Squamous cell carcinoma)



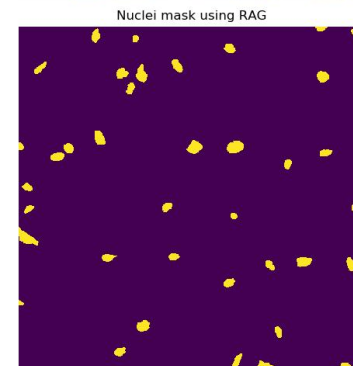
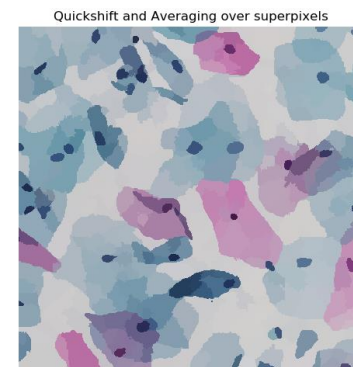
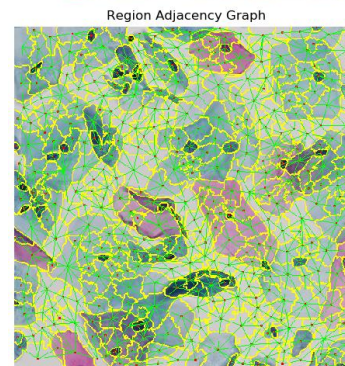
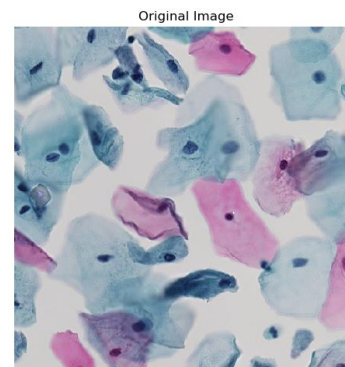
# Cell Data Generation

---

- **We employ two methods:**
  - **Graph based cell detection**
  - **Patch based data generation**

# Graph-based Cell Detection

- Superpixel generation using QuickShift<sup>Δ</sup> technique
- Averaging pixel intensities over superpixel regions
- Create a Graph:
  - Node = Centroid of superpixel regions
  - Edge = Connecting adjacent nodes
  - Edge weight =  $\left| \sqrt{n_1^2 - n_2^2} \right|$
- Graph Cut, with threshold = 59 (chosen empirically)
- Output: Nuclei mask

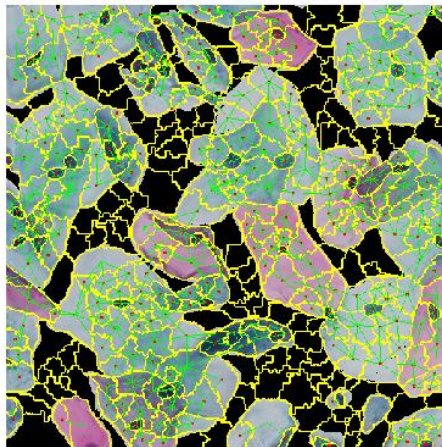


<sup>Δ</sup>A. Vedaldi and S. Soatto. Quick shift and kernel methods for mode seeking. In Proc. ECCV, 2008.

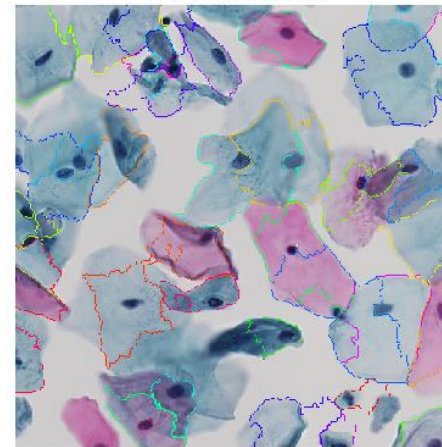
# Graph-based Cell Detection

- ▶ **Cell Detection**
  - ▶ **Threshold to remove background**
  - ▶ **Join regions - nodes adjacent to nuclei node**
  - ▶ **Working on the problem to improve the cell detection**

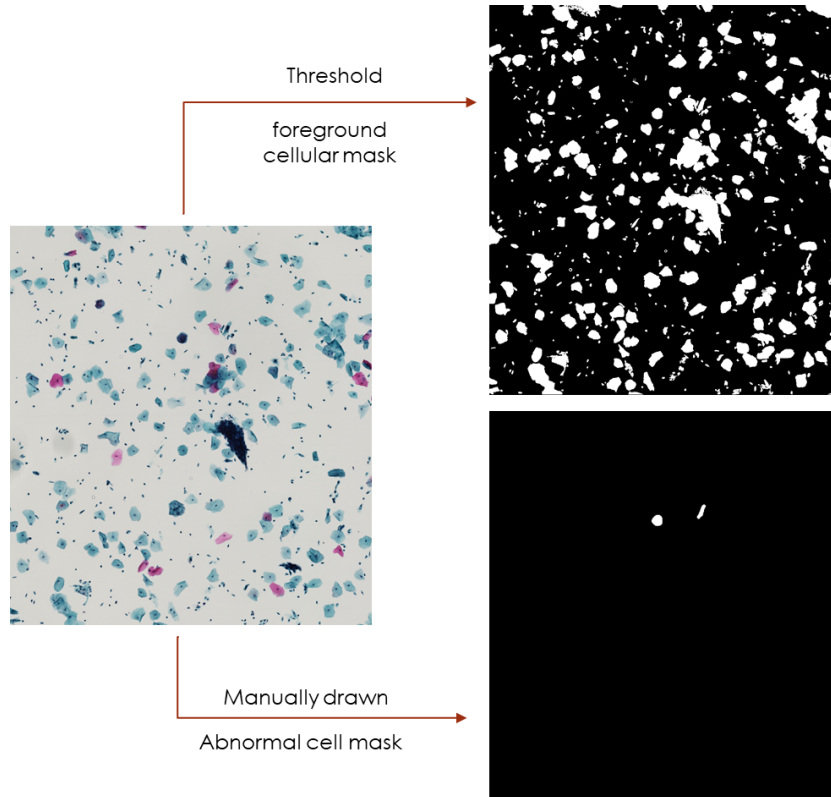
Graph structure connecting cellular regions



Cell segmentation



# Patch-Based Data Generation



- ▶ 128x128 patch generation from sliding window technique (stride = 64)
- ▶ Patches containing more than 75% of background were discarded.
- ▶ Ground truth labels for each patch were generated w.r.t abnormal cell mask.
  - ▶ Object area > 20% Patch area → Abnormal
  - ▶ Otherwise → Normal

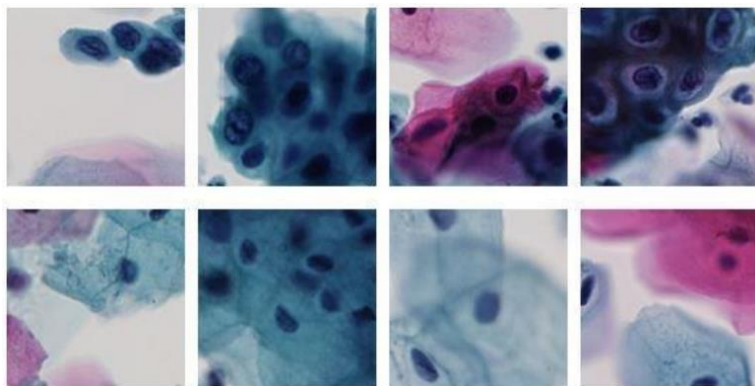
# Cell Data for Classification

## ➤ Patch Based Data Generation

- 4120 multi-cell images
  - Abnormal images: 2060
  - Normal images: 2060

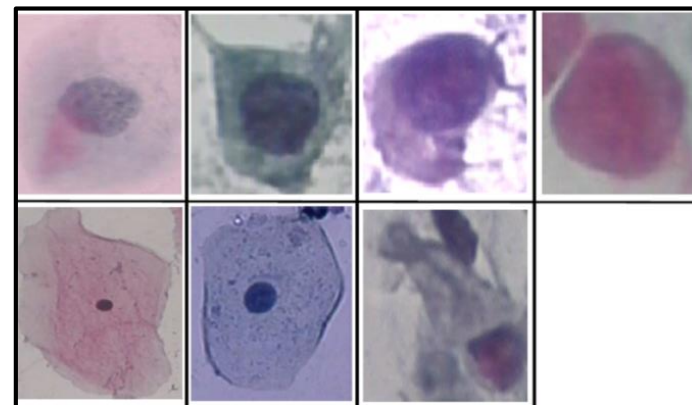
## ➤ Herlev Pap Smear Dataset

- 917 individual cell images
  - Abnormal cells: 675
  - Normal cells: 242



Abnormal

Normal





# Classification

## ► CNN-based Classifier

- Fine-tuning models initialized with pre-trained ImageNet weights.
- No. of Epochs = 500
- Batch Size = 32
- Optimizer: Stochastic Gradient Descent
- Learning rate = 0.005, Momentum = 0.9
- Loss Function: Cross Entropy Loss
- Trained weights saved at epoch with best accuracy

Input for CNN Classifier		Total	Patch data	Herlev data
Training	Normal	1396	1200	196
	Abnormal	1760	1200	560
Validation	Normal	246	200	46
	Abnormal	315	200	115
Testing	Normal	660	660	-
	Abnormal	660	660	-
Total		5037	4120	917

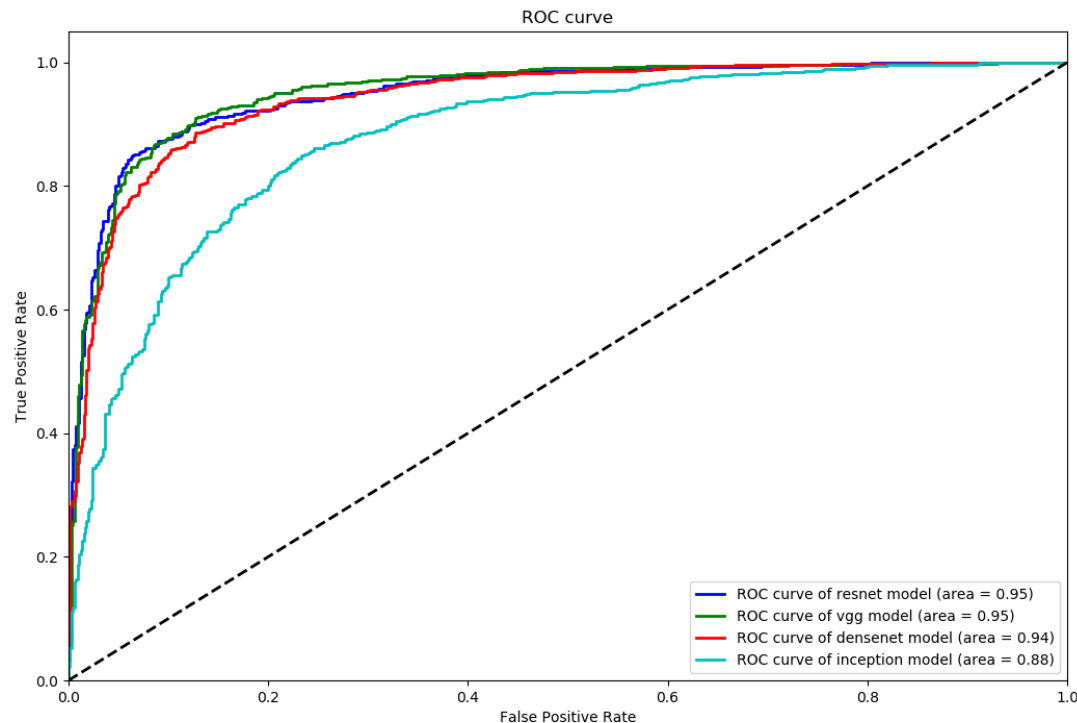
# Classification Results

- PyTorch Deep Learning Platform.
- Models run on Nvidia DGX-1.
- VGG-19 was observed to give better performance

Model	Confusion matrix $\begin{bmatrix} TN & FP \\ FN & TP \end{bmatrix}$	Accuracy	Precision	Recall	F1-score	ROC_AUC
Resnet-50	$\begin{bmatrix} 589 & 71 \\ 78 & 582 \end{bmatrix}$	0.8871	0.8913	0.8818	0.8865	0.95
<b>VGG-19</b>	$\begin{bmatrix} \mathbf{581} & \mathbf{79} \\ \mathbf{68} & \mathbf{592} \end{bmatrix}$	<b>0.8886</b>	<b>0.8823</b>	<b>0.8970</b>	<b>0.8896</b>	<b>0.95</b>
Densenet-121	$\begin{bmatrix} 611 & 49 \\ 131 & 529 \end{bmatrix}$	0.8636	0.9152	0.8015	0.8546	0.94
Inception_v3	$\begin{bmatrix} 429 & 231 \\ 57 & 603 \end{bmatrix}$	0.7818	0.7230	0.9136	0.8072	0.88

# Classification Results

- PyTorch Deep Learning Platform.
- Models run on Nvidia DGX-1.
- VGG-19 was observed to give better performance
- However, VGG-19 would take longer time to train and is computationally expensive compared to ResNet-50.



# Conclusion

- This is an opportunity to explore use of Machine Learning and Artificial Intelligence for improving efficiency in cervical cytology.
- Cell detection and classification pose great challenges due to
  - Limited availability of proper annotated data
  - Complexity due to overlapping cells
- Our method performs well as a binary classifier.
- Classification task could be extended to perform multi-class classification.
- Need to develop robust cell detection algorithm using Reinforcement learning techniques



U.S. National Library  
of Medicine



# Thank you!

Email me at:

[ssbw5@mst.edu](mailto:ssbw5@mst.edu)

