Pathology Image Analysis
Workshop

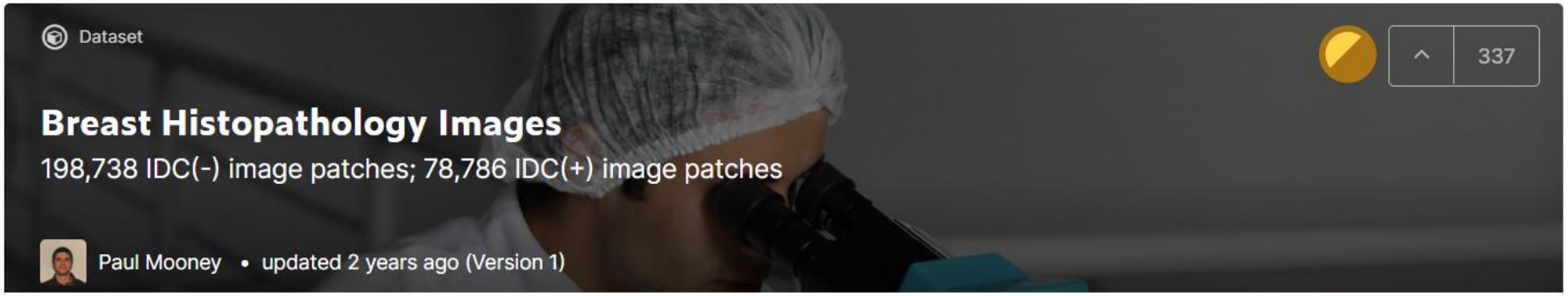
Albert C. Yang, M.D., Ph.D.

Institutes of Brain Science, National Yang-Ming University

May 28, 2020

accyang@gmail.com

Breast Histopathology Dataset



Data Tasks Kernels (32) Discussion (6) Activity Metadata Download (1 GB) New Notebook

Usability 7.5 License CC0: Public Domain Tags online image galleries, health, image data, cancer, womens health and 3 more

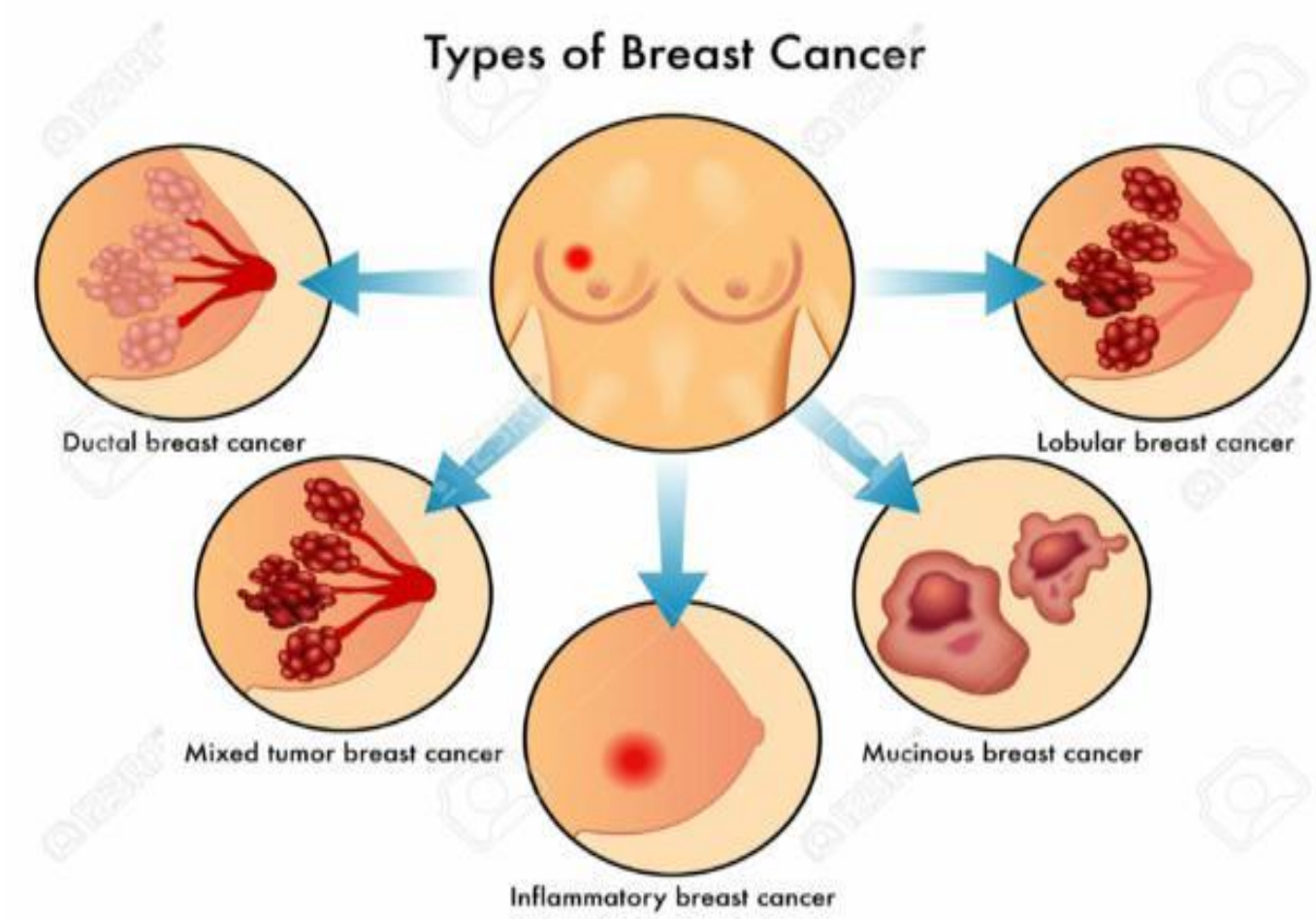
Description

Context

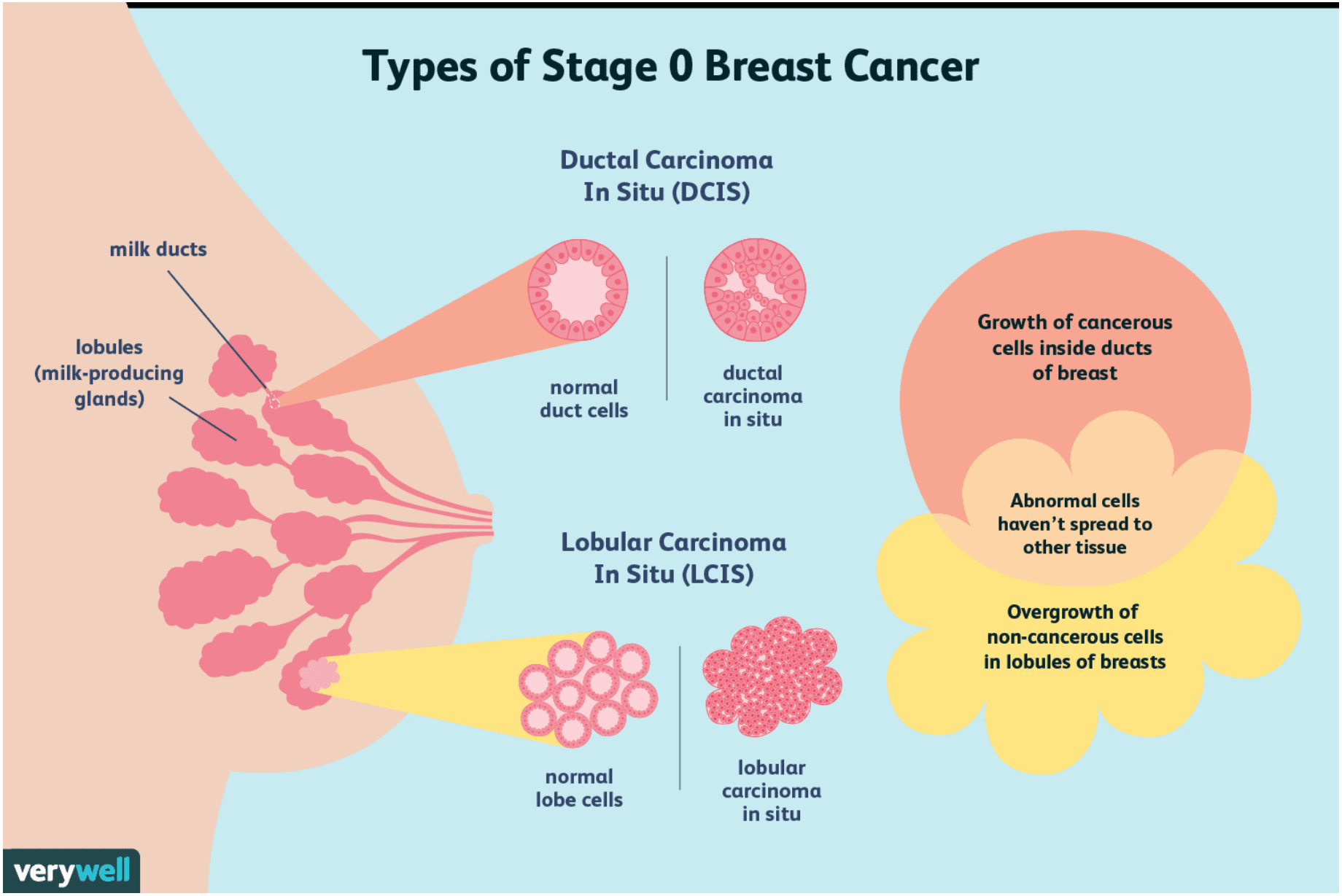
Invasive Ductal Carcinoma (IDC) is the most common subtype of all breast cancers. To assign an aggressiveness grade to a whole mount sample, pathologists typically focus on the regions which contain the IDC. As a result, one of the common pre-processing steps for automatic aggressiveness grading is to delineate the exact regions of IDC inside of a whole mount slide.

https://www.kaggle.com/paultimothymooney/breast-histopathology-images

Types of Breast Cancer



https://peekerhealth.com/know-more-about-types-of-breast-cancer/



https://www.verywellhealth.com/breast-cancer-staging-stage-zero-429887

BREAST CANCER STAGES

BREAST CANCER STAGING MEASURES THE SPREAD OF THE DISEASE UPON DIAGNOSIS.
IN ORDER TO DETERMINE THE CHOICE OF TREATMENT, IT IS VERY IMPORTANT TO STAGE THE CANCER

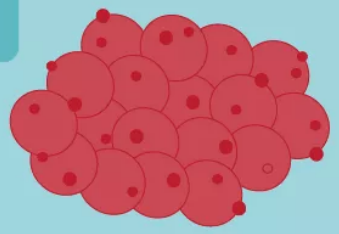
STAGE	0	1	2	3	4
TUMOR SIZE	VERY SMALL, INSIDE THE GLANDS	LESS THAN 2 CM	5-2 CM	5 CM AND LARGER	ANY SIZE
LYMPH NODES	NO CANCER	NO CANCER	AFFECTED BY CANCER	AFFECTED BY CANCER; CANCER HAS REACHED THE MUSCLES AND SKIN	AFFECTED BY CANCER
SPREADING	CONFINED TO THE BREAST AREA, NOT OUTSIDE	CONFINED TO THE BREAST AREA, NOT OUTSIDE	CONFINED TO THE BREAST AREA, NOT OUTSIDE	CONFINED TO THE BREAST AREA, NOT OUTSIDE	CANCER HAS SPREAD OUTSIDE THE BREAST AREA TO ANY PART OF THE BODY
EV	-EV	-EV	-EV	-EV	+++EV
5 YEAR SURVIVAL RATE	100%	100%	87%	61%	20%

+962 6 5530800 +962 6 5530870 JBCPjordan Jordan Breast Cancer Program www.jbcp.jo

https://medcaretips.com/staging-of-breast-cancer/

TNM System for Staging Breast Cancer

T



Tumor size

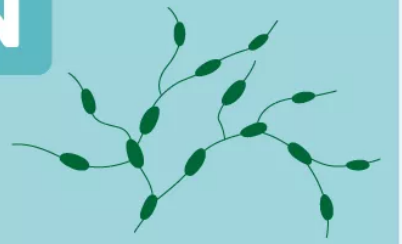
T-1: 0-2 centimeters

T-2: 2-5 centimeters

T-3: >5 centimeters

T-4: Tumor has broken through skin or attached to chest wall

N



Lymph Node Status

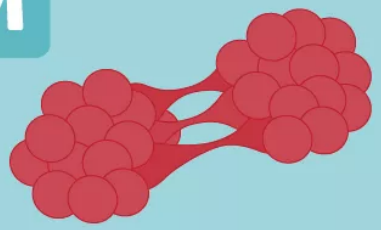
N-0: Surgeon can't feel any nodes

N-1: Surgeon can feel swollen nodes

N-2: Nodes feel swollen and lumpy

N-3: Swollen nodes located near collarbone

M



Metastasis

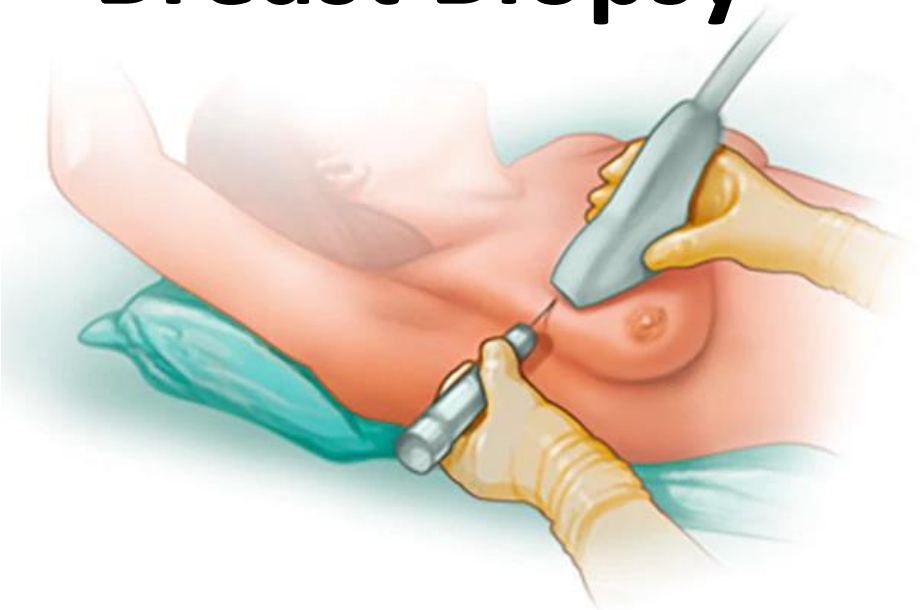
M-0: Tested nodes are cancer-free

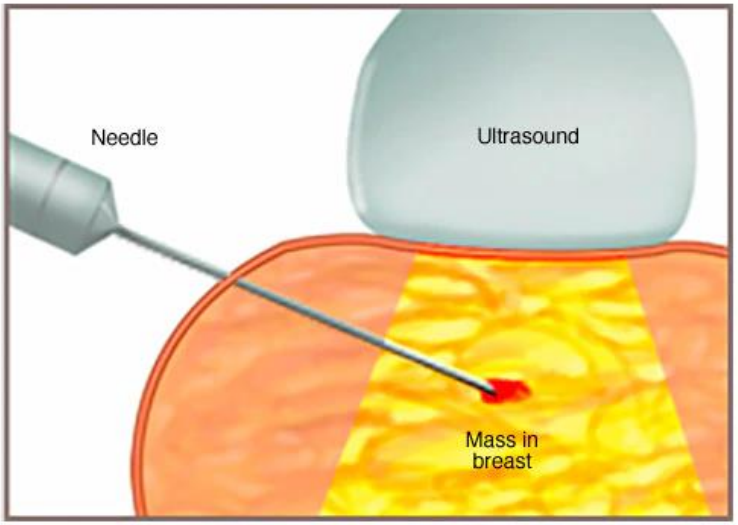
M-1: Tested nodes show cancer cells or micrometastasis

verywell

https://www.verywellhealth.com/know-your-breast-tumor-size-4114640

Breast Biopsy





https://www.mayoclinic.org/tests-procedures/breast-biopsy/about/pac-20384812

Breast Histopathology Dataset

- Breast cancer is the most common form of cancer in women.
- Invasive ductal carcinoma (IDC) is the most common form of breast cancer.
- Accurately identifying and categorizing breast cancer subtypes is an important clinical task, and automated methods can be used to save time and reduce error.

Breast Histopathology Dataset

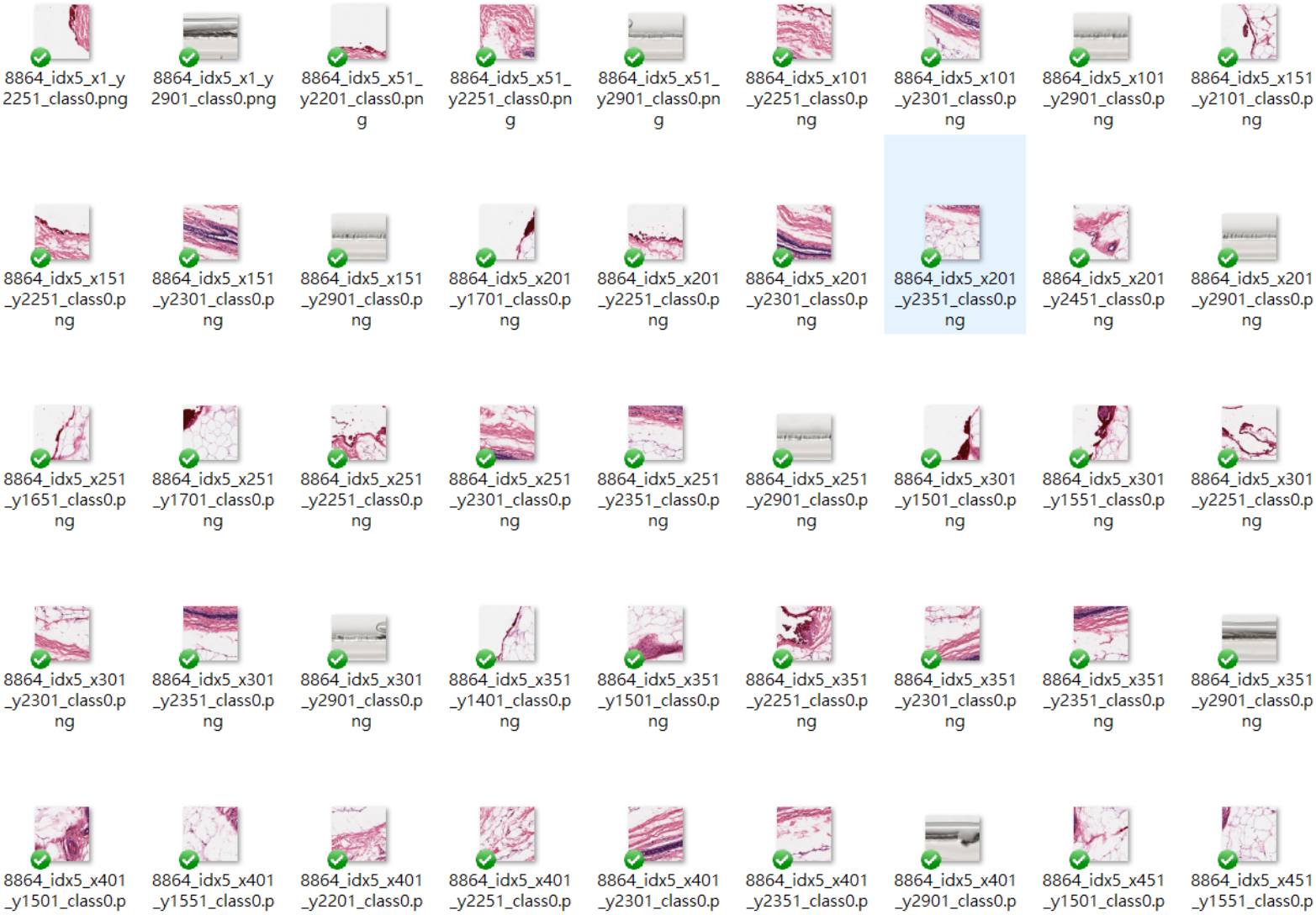
- The original dataset consisted of 162 whole mount slide images of Breast Cancer (BCa) specimens scanned at 40x.

- From that, 277,524 patches of size 50 x 50 were extracted (198,738 IDC negative and 78,786 IDC positive).

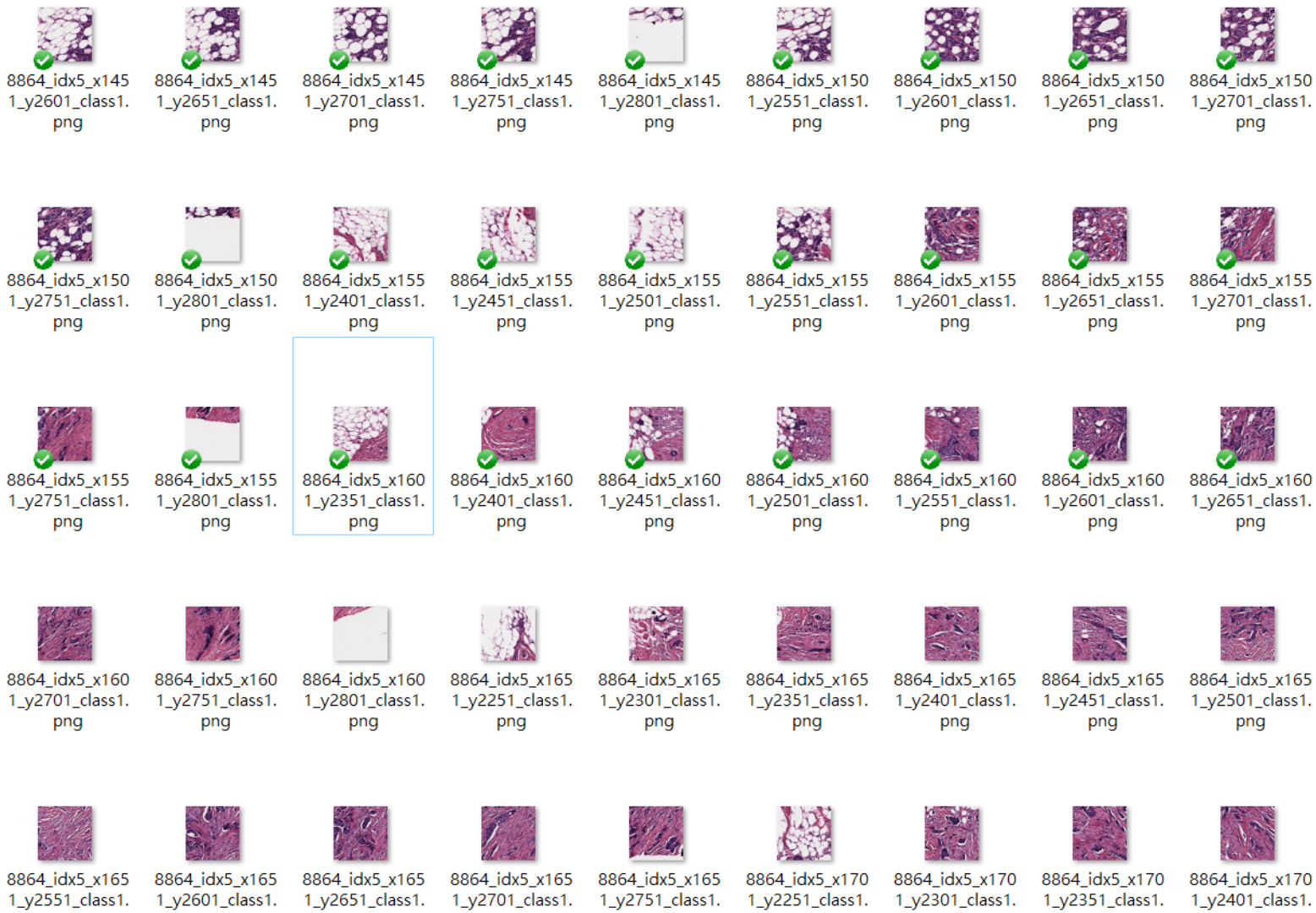
- Each patch's file name is of the format: uxXyYclassC.png
  - example 10253idx5x1351y1101class0.png . Where u is the patient ID (10253idx5), X is the x-coordinate of where this patch was cropped from, Y is the y-coordinate of where this patch was cropped from, and C indicates the class where 0 is non-IDC and 1 is ID

- **Simplified 1000 vs. 1000 breast histopathology image data**

Normal Part of Tissue



Invasive Ductal Carcinoma



Breast Histopathology Dataset

..			本機磁碟		
Breast Histopathology			檔案資料夾	2020/5/28 ...	
cnn.m	2,324	885	M 檔案	2020/5/28 ...	B90D2...
confusionmatStats.m	2,548	833	M 檔案	2015/10/1...	7F3D6...
image2datastore.m	462	269	M 檔案	2019/10/1...	9F9FE...

Labdata > Breast Histopathology >

名稱	修改日期	類型	大小
0	2020/5/28 上午 08:09	檔案資料夾	
1	2020/5/28 上午 08:35	檔案資料夾	
Label.txt	2020/5/28 上午 10:02	文字文件	1 KB

Label.txt - 記事本

檔案(F) 編輯(E) 格式(O) 檢視(V) 說明

```
0: non-tumor parts of tissue image
1: invasive ductal carcinoma
```

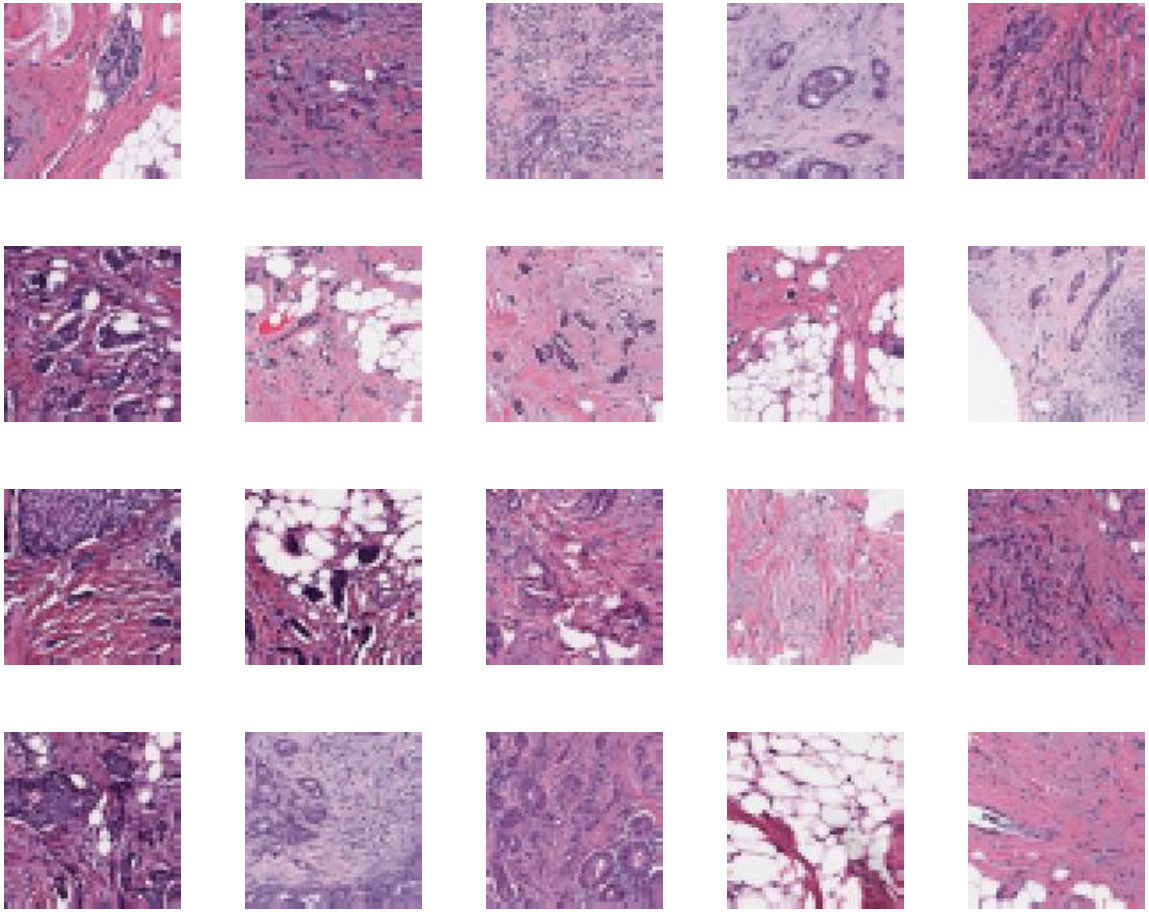
Read Image Data

- img = imread('8864_idx5_x1_y2251_class0.png');
- imshow(img)

Show 20 Random Images

```
files = dir('*.png');
figure;
perm = randperm(1000,20);
for i = 1:20
    subplot(4,5,i);
    img = imread(files(perm(i)).name);
    imshow(img);
end
```

Show 20 Random Images

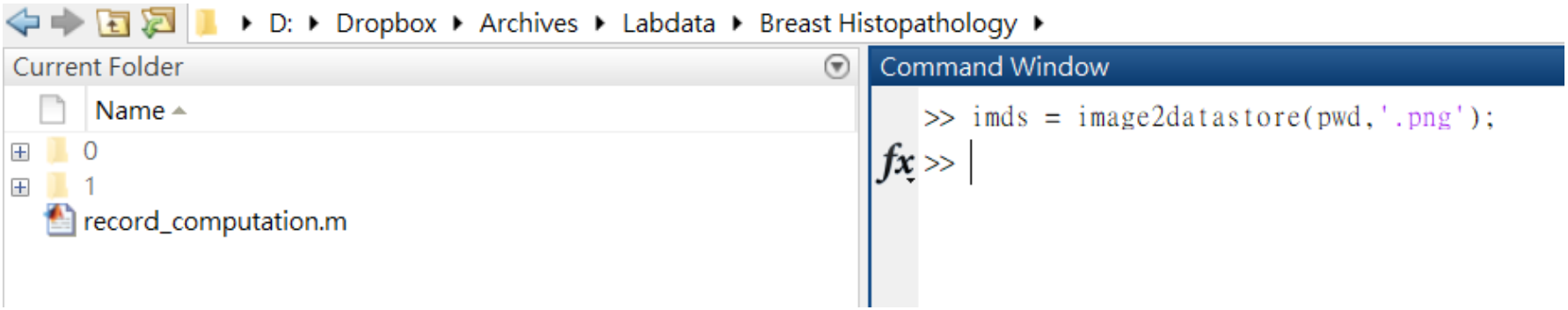


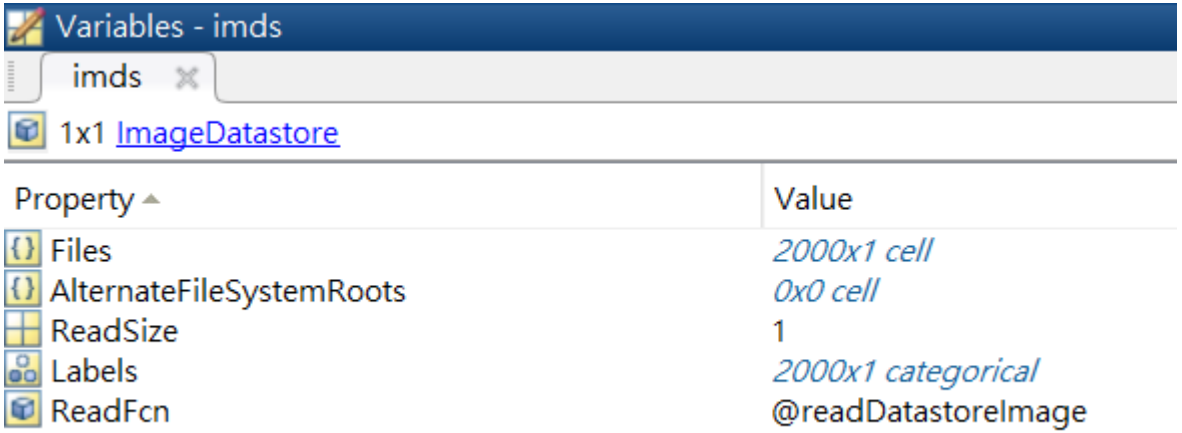
Create Image Datastore

```
function imds = image2datastore(datapath,file_ext)

% Get folder list
dinfo = dir();
dirFlags = [dinfo.isdir];
dinfo = dinfo(dirFlags);
dinfo(ismember( {dinfo.name}, {'.', '..'})) = [];

% Create image datastore using foldername and input file extension
filelocation = {};
for i=1:length(dinfo)
    filelocation{i} = [datapath '\' dinfo(i).name];
end
imds = imageDatastore(filelocation,'FileExtensions',file_ext,'LabelSource','foldernames');

end
```

Create Image Datastore

- imds = image2datastore(pwd,'.png');





Count Number of Images for Each Label

- labelCount = countEachLabel(imds);
- labelCount = labelCount.Count;
- min_labelCount = min(labelCount);

Workspace

Name	Value
imds	*1x1 ImageDatastore*
labelCount	[1000;1000]

Specify Image Size

```
img = readimage(imds,1);
imgsize = size(img);
if length(imgsize)==2
    imgsize(3) = 1;
end
```

Specify Training and Validation Sets

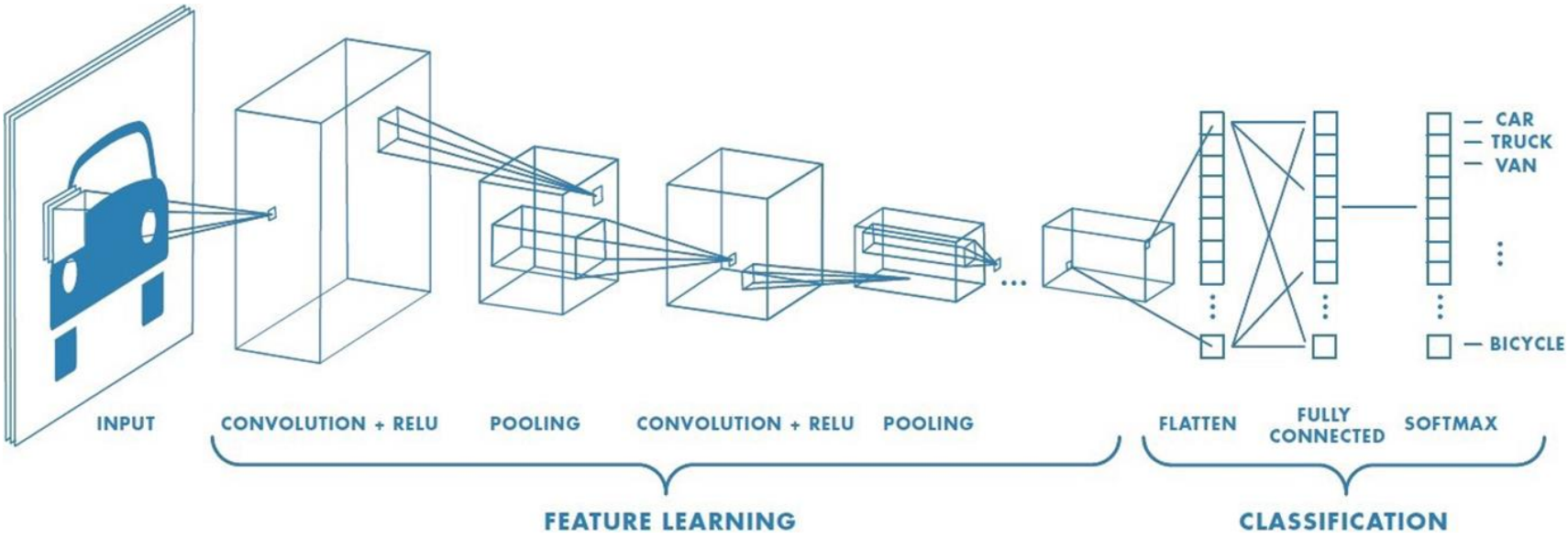
- train_ratio = 0.7;
- numTrainFiles = fix(min_labelCount*train_ratio);
- [imdsTrain,imdsValidation] = splitEachLabel(imds,numTrainFiles,'randomize');

Workspace

Name	Value
imds	*1x1 ImageDatastore*
imdsTrain	*1x1 ImageDatastore*
imdsValidation	*1x1 ImageDatastore*
img	*50x50x3 uint8*
imgsize	[50,50,3]
labelCount	[1000;1000]
min_labelCount	1000
numTrainFiles	700
train_ratio	0.7000

Specify Convolution Layer Parameters

- filter_size = 3;
- num_filters = 8;



Specify CNN Architechure

```
layers = [
    imageInputLayer(imgsize)

    convolution2dLayer(filter_size,num_filters,'Padding','same')
    batchNormalizationLayer
    reluLayer

    maxPooling2dLayer(2,'Stride',2)

    convolution2dLayer(filter_size,num_filters*2,'Padding','same')
    batchNormalizationLayer
    reluLayer

    maxPooling2dLayer(2,'Stride',2)

    convolution2dLayer(filter_size,num_filters*4,'Padding','same')
    batchNormalizationLayer
    reluLayer

    fullyConnectedLayer(length(labelCount))
    softmaxLayer
    classificationLayer];
```

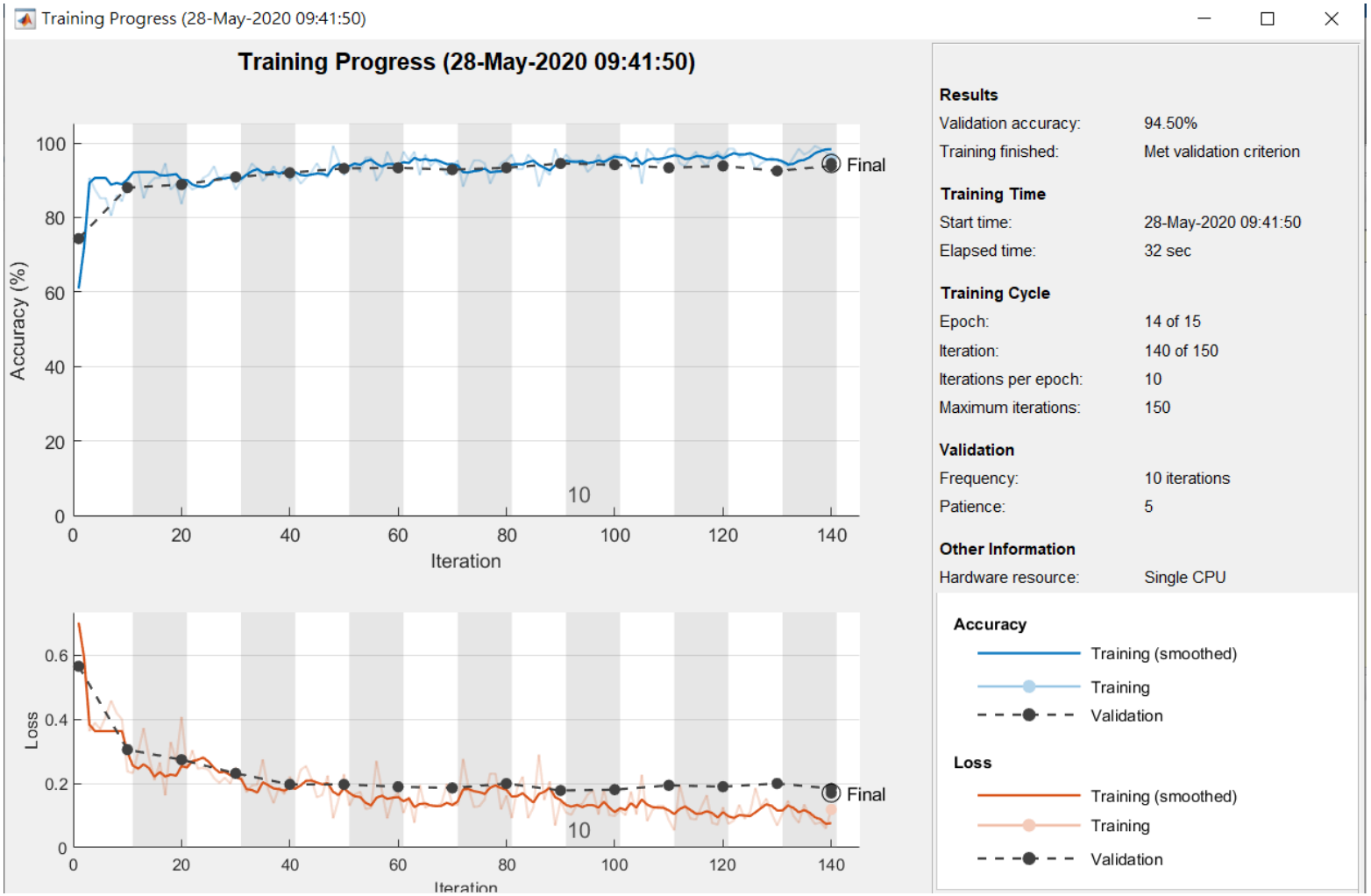
Specify Training Options

```
options = trainingOptions('sgdm', ...
    'InitialLearnRate',0.001, ...
    'MaxEpochs',15, ...
    'Shuffle','every-epoch', ...
    'ValidationData',imdsValidation, ...
    'ValidationFrequency',10, ...
    'Verbose',false, ...
    'Plots', 'training-progress');
```

Start Training

- tic;
- [net netinfo]= trainNetwork(imdsTrain,layers,options);
- toc;

Training Progress

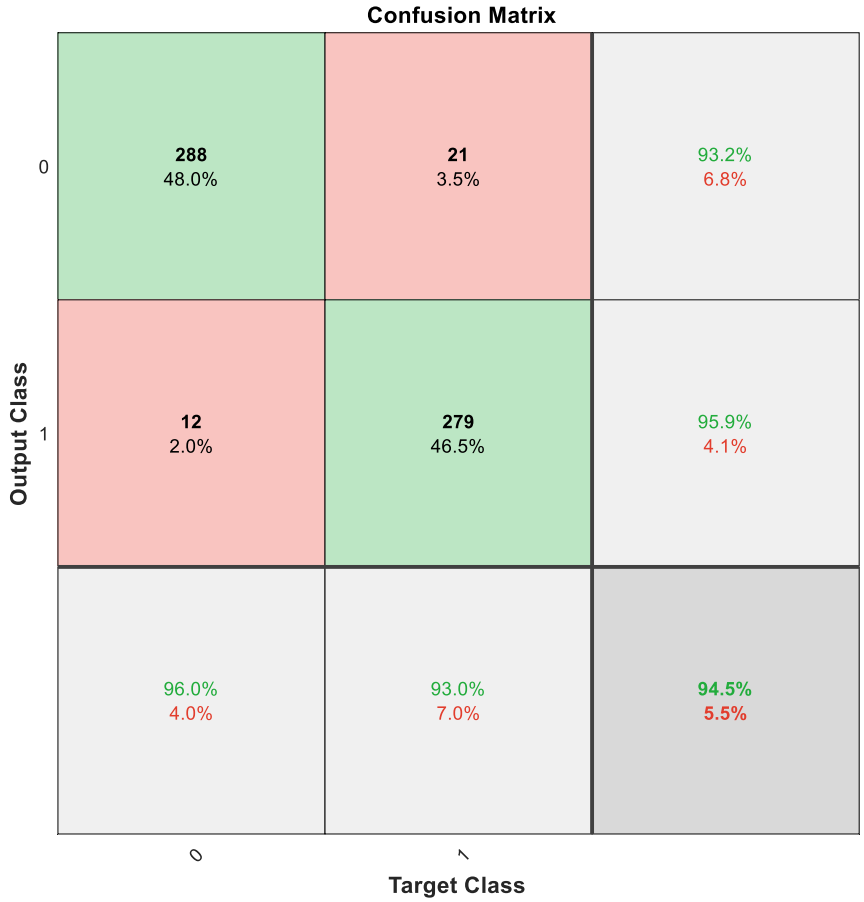


Compute Accuracy

- YPred = classify(net,imdsValidation);
- YValidation = imdsValidation.Labels;
-
- accuracy = sum(YPred == YValidation)/numel(YValidation);

Plot Confusion Matrix

- plotconfusion(YValidation,YPred)



A CNN Code

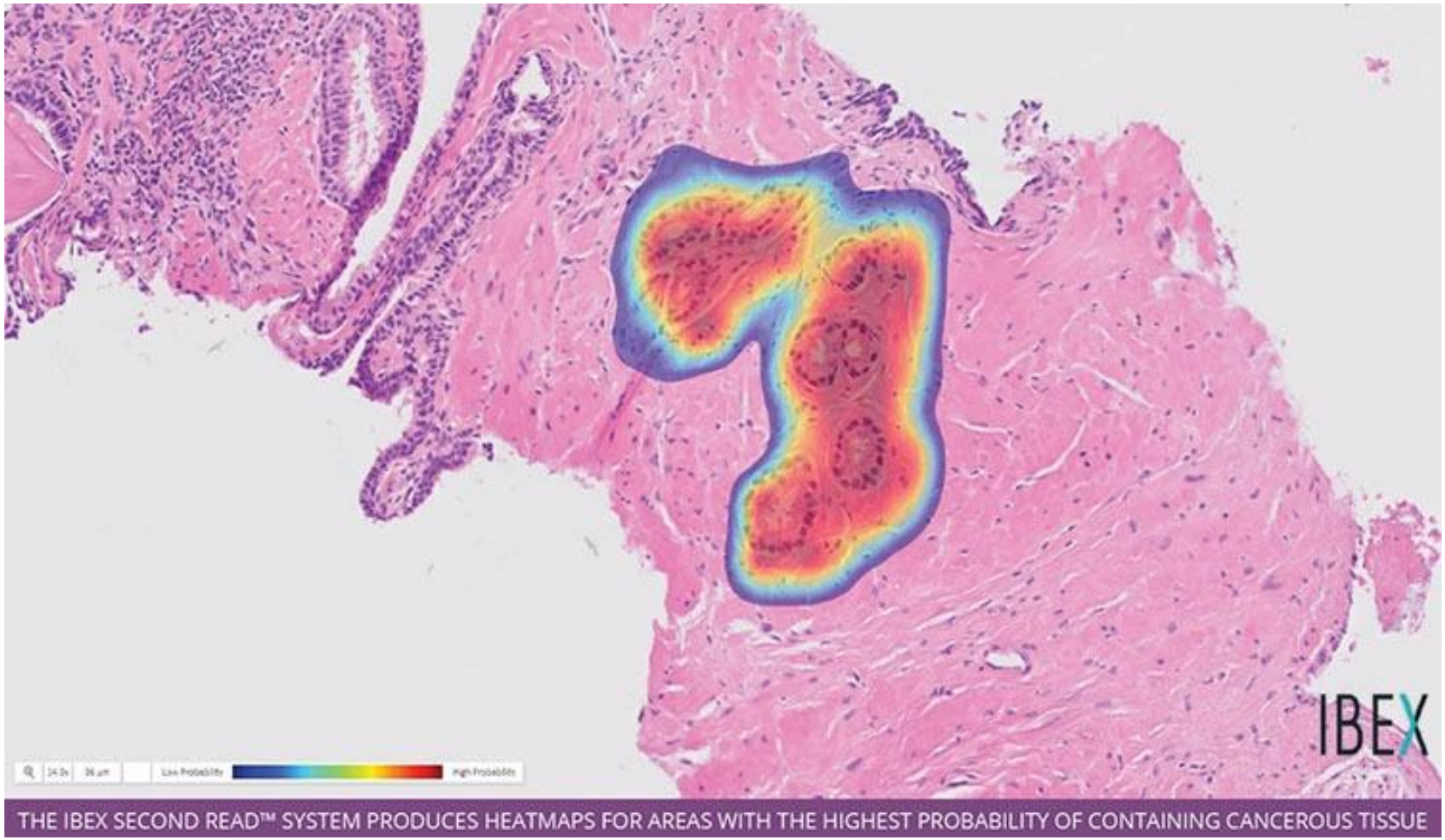
cnn.m | record_computation.m

```
% cnn.m
% Dr. Albert C. Yang, MD, PhD. 2019/10/17
% Laboratory of Precision Psychiatry (http://www.precisionpsychiatry.org)

function [net netinfo nstats] = cnn(imds,train_ratio,show_progress,filter_size,num_filters)

if nargin<3 || isempty(show_progress)==1
    show_progress = 1;
    plots_option = 'training-progress';
end

if nargin<4 || isempty(filter_size)==1
    filter_size = 4;
    plots_option = 'training-progress';
end

if nargin<5 || isempty(num_filters)==1
    num_filters = 8;
    plots_option = 'training-progress';
```

Run CNN at Once…

- imds = image2datastore(pwd,'.png');
- [net netinfo nstats] = cnn(imds,0.7,1,3,8);

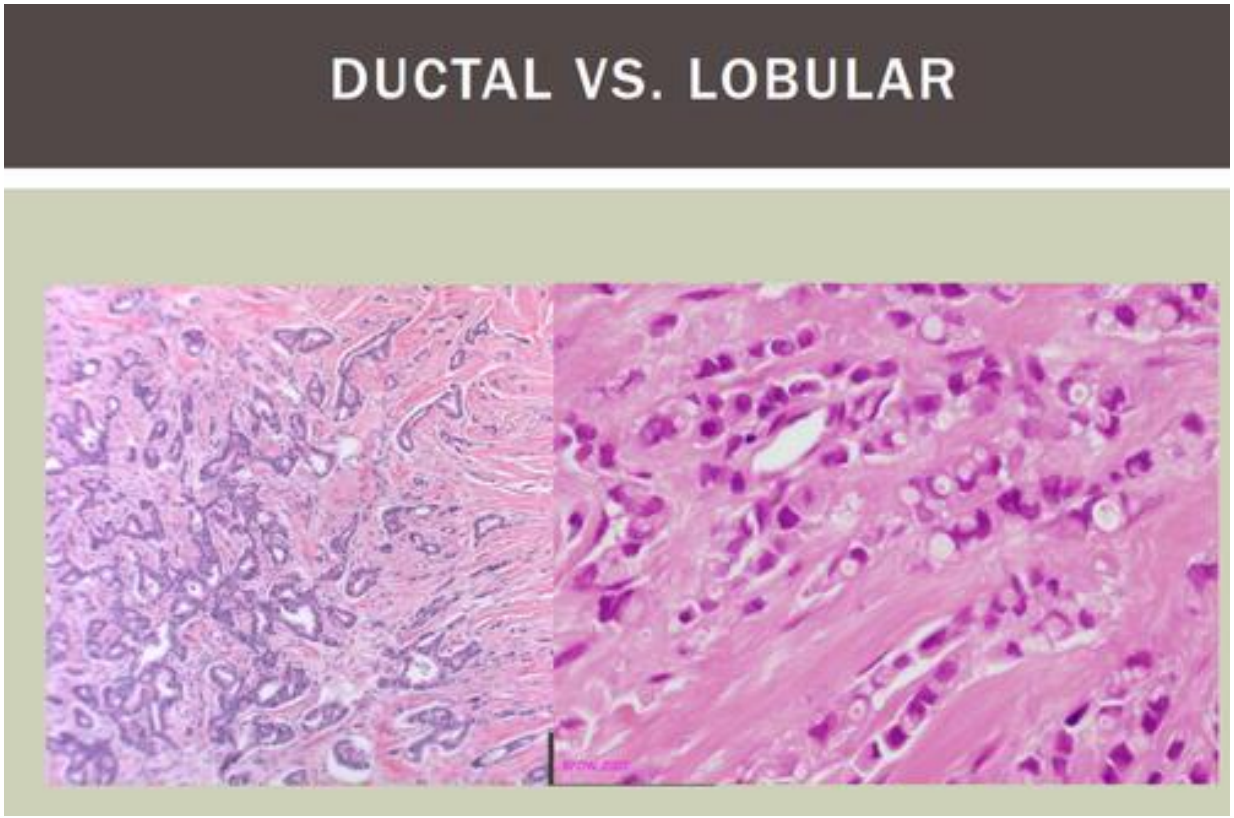
Challenges - Heatmap



https://www.nanalyze.com/2019/10/ai-cancer-diagnostics/

Challenge - Multilabel Classification

- Ductal vs. Lobular Type of Breast Cancer



https://quizlet.com/206900783/overview-of-breast-cancer-breast-pathology-1-and-2-flash-cards/