



Pathology Image Analysis Workshop

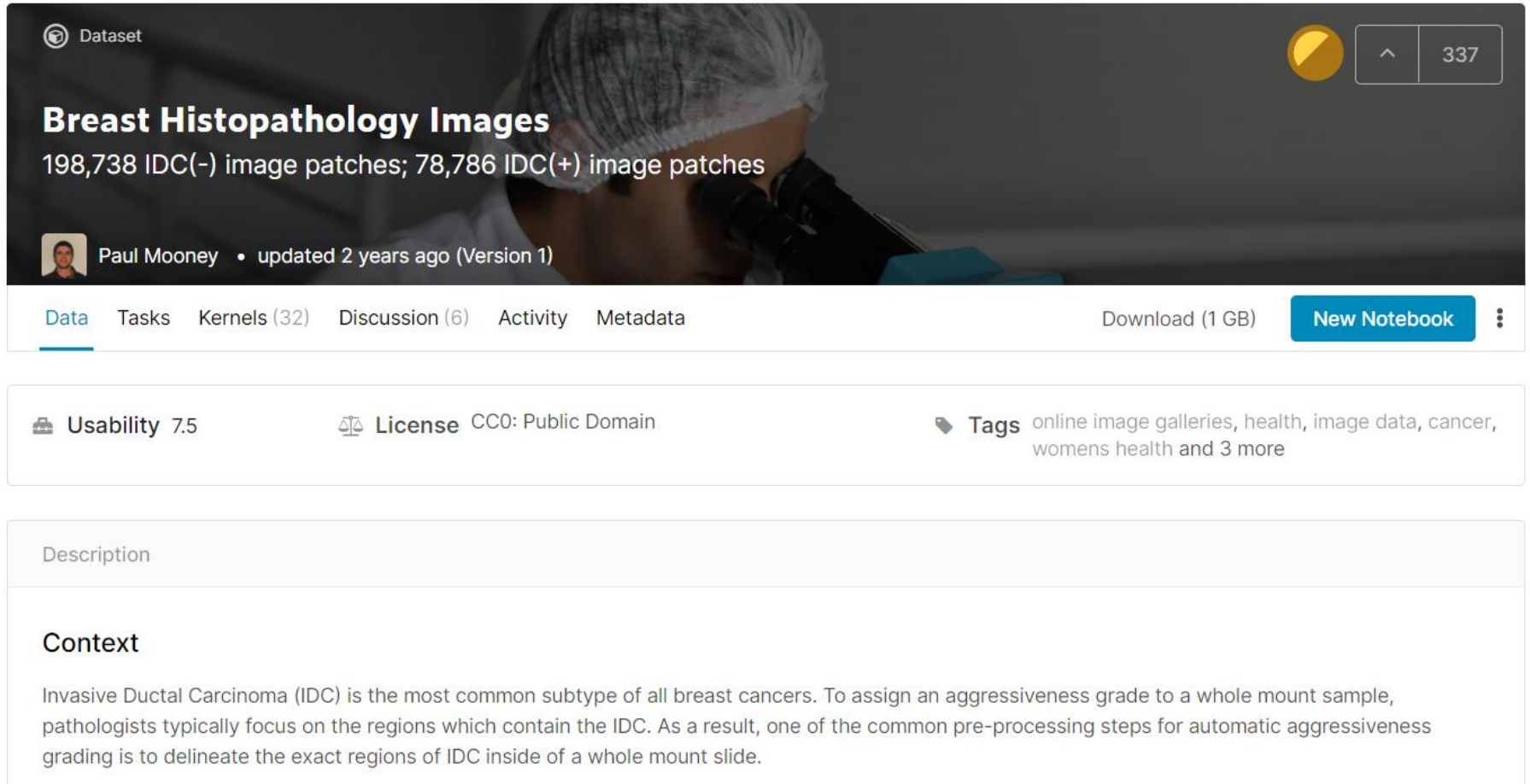
Albert C. Yang, M.D., Ph.D.

Institutes of Brain Science, National Yang-Ming University

May 28, 2020

accyang@gmail.com

Breast Histopathology Dataset



The screenshot shows the Kaggle dataset page for 'Breast Histopathology Images'. The header includes a 'Dataset' label, a gold badge, and a view count of 337. The main title is 'Breast Histopathology Images' with a subtitle '198,738 IDC(-) image patches; 78,786 IDC(+) image patches'. The creator is Paul Mooney, updated 2 years ago (Version 1). Navigation tabs include Data, Tasks, Kernels (32), Discussion (6), Activity, and Metadata. A 'Download (1 GB)' link and a 'New Notebook' button are present. Below the navigation are sections for Usability (7.5), License (CC0: Public Domain), and Tags (online image galleries, health, image data, cancer, womens health and 3 more). The 'Description' section is partially visible, showing a 'Context' heading and the start of a paragraph about Invasive Ductal Carcinoma (IDC).

Dataset

Breast Histopathology Images

198,738 IDC(-) image patches; 78,786 IDC(+) image patches

Paul Mooney • updated 2 years ago (Version 1)

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

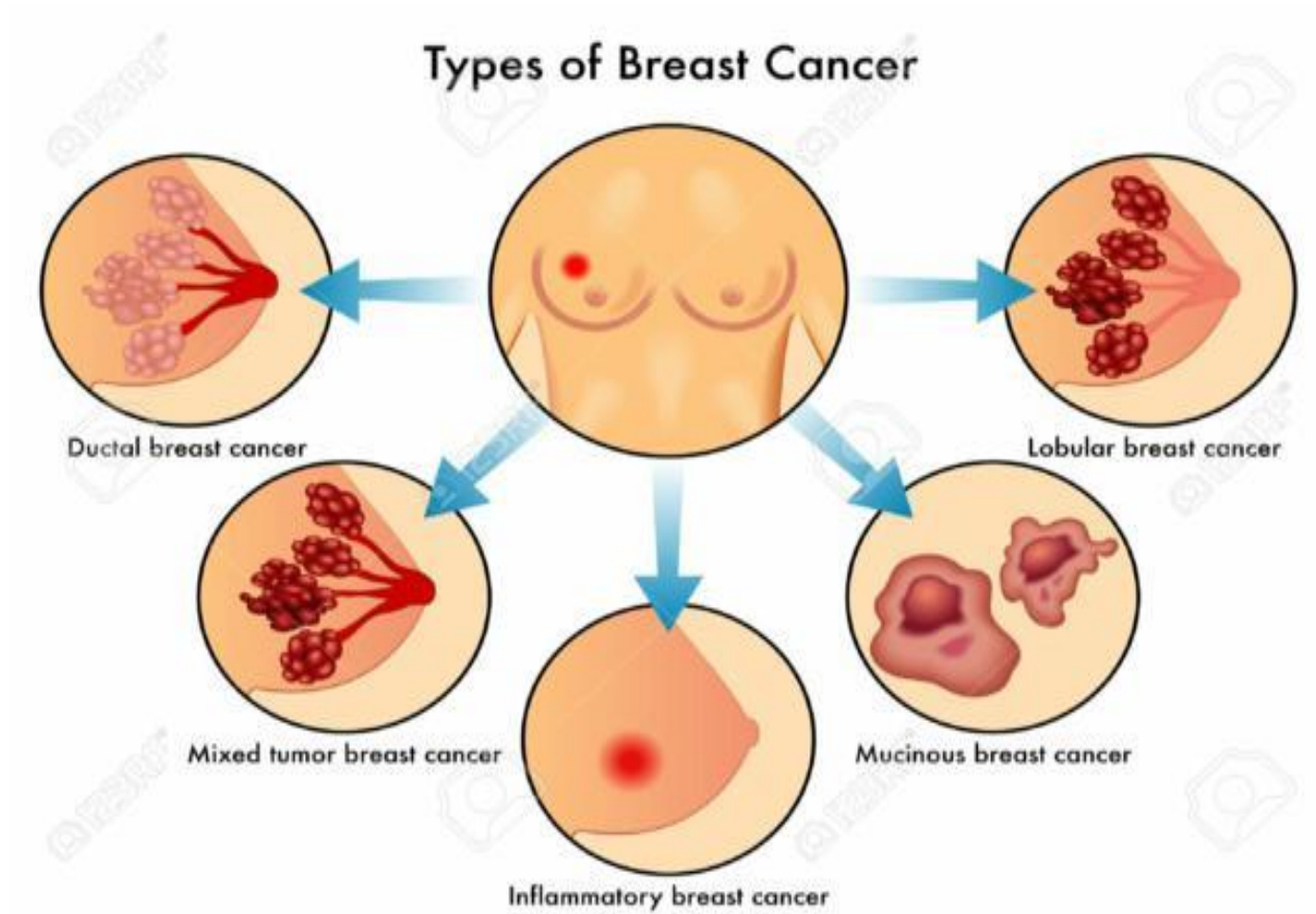
Description

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/>

Types of Stage 0 Breast Cancer

Ductal Carcinoma In Situ (DCIS)

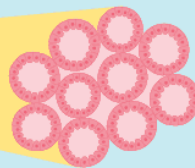


normal duct cells

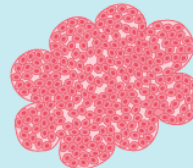


ductal carcinoma in situ

Lobular Carcinoma In Situ (LCIS)



normal lobe cells



lobular carcinoma in situ

Growth of cancerous cells inside ducts of breast

Abnormal cells haven't spread to other tissue

Overgrowth of non-cancerous cells in lobules of breasts

milk ducts
lobules (milk-producing glands)

BREAST CANCER STAGES

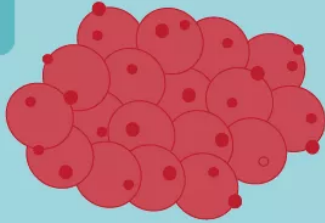
BREAST CANCER STAGING MEASURES THE SPREAD OF THE DISEASE UPON DIAGNOSIS. IN ORDER TO DETERMINE THE CHOICE OF TREATMENT, IT IS VERY IMPORTANT TO STAGE THE CANCER



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%

TNM System for Staging Breast Cancer

T



Tumor size

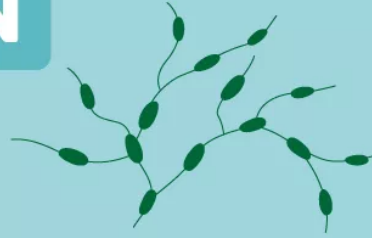
T-1: 0-2 centimeters

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

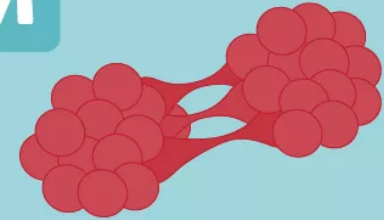
N-0: Surgeon can't feel any nodes

N-1: Surgeon can feel swollen nodes

N-2: Nodes feel swollen and lumpy

N-3: Swollen nodes located near collarbone

M

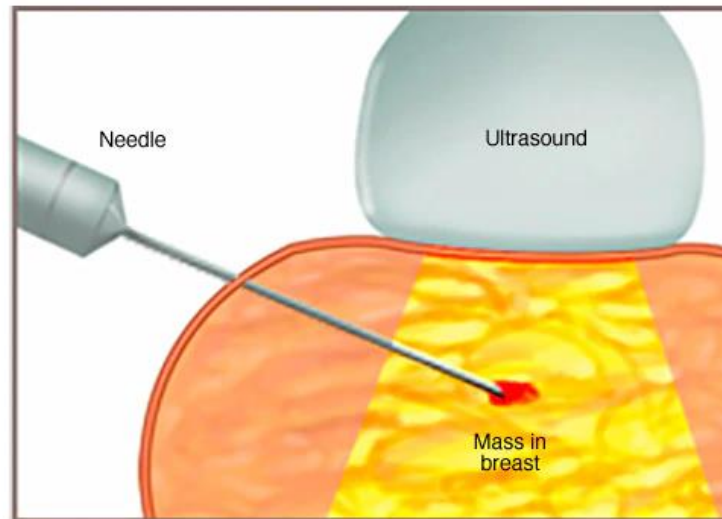
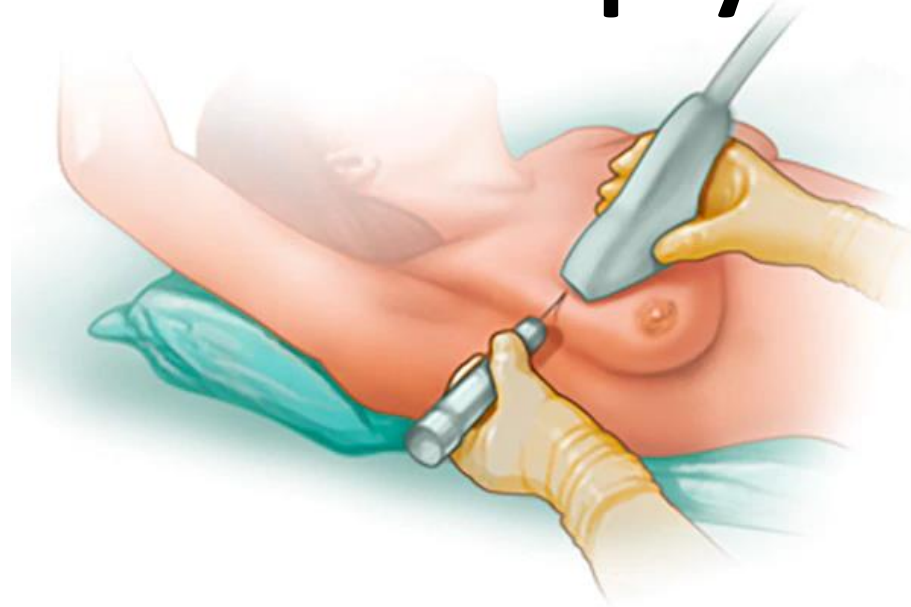


Metastasis

M-0: Tested nodes are cancer-free

M-1: Tested nodes show cancer cells or micrometastasis

Breast Biopsy



© MAYO FOUNDATION FOR MEDICAL EDUCATION AND RESEARCH. ALL RIGHTS RESERVED.

<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



8864_idx5_x1_y2251_class0.png



8864_idx5_x1_y2901_class0.png



8864_idx5_x51_y2201_class0.png



8864_idx5_x51_y2251_class0.png



8864_idx5_x51_y2901_class0.png



8864_idx5_x101_y2251_class0.png



8864_idx5_x101_y2301_class0.png



8864_idx5_x101_y2901_class0.png



8864_idx5_x151_y2101_class0.png



8864_idx5_x151_y2251_class0.png



8864_idx5_x151_y2301_class0.png



8864_idx5_x151_y2901_class0.png



8864_idx5_x201_y1701_class0.png



8864_idx5_x201_y2251_class0.png



8864_idx5_x201_y2301_class0.png



8864_idx5_x201_y2351_class0.png



8864_idx5_x201_y2451_class0.png



8864_idx5_x201_y2901_class0.png



8864_idx5_x251_y1651_class0.png



8864_idx5_x251_y1701_class0.png



8864_idx5_x251_y2251_class0.png



8864_idx5_x251_y2301_class0.png



8864_idx5_x251_y2351_class0.png



8864_idx5_x251_y2901_class0.png



8864_idx5_x301_y1501_class0.png



8864_idx5_x301_y1551_class0.png



8864_idx5_x301_y2251_class0.png



8864_idx5_x301_y2301_class0.png



8864_idx5_x301_y2351_class0.png



8864_idx5_x301_y2901_class0.png



8864_idx5_x351_y1401_class0.png



8864_idx5_x351_y1501_class0.png



8864_idx5_x351_y2251_class0.png



8864_idx5_x351_y2301_class0.png



8864_idx5_x351_y2351_class0.png



8864_idx5_x351_y2901_class0.png



8864_idx5_x401_y1501_class0.png



8864_idx5_x401_y1551_class0.png



8864_idx5_x401_y2201_class0.png



8864_idx5_x401_y2251_class0.png



8864_idx5_x401_y2301_class0.png



8864_idx5_x401_y2351_class0.png



8864_idx5_x401_y2901_class0.png



8864_idx5_x451_y1501_class0.png



8864_idx5_x451_y1551_class0.png

Invasive Ductal Carcinoma



8864_idx5_x145
1_y2601_class1.
png



8864_idx5_x145
1_y2651_class1.
png



8864_idx5_x145
1_y2701_class1.
png



8864_idx5_x145
1_y2751_class1.
png



8864_idx5_x145
1_y2801_class1.
png



8864_idx5_x150
1_y2551_class1.
png



8864_idx5_x150
1_y2601_class1.
png



8864_idx5_x150
1_y2651_class1.
png



8864_idx5_x150
1_y2701_class1.
png



8864_idx5_x150
1_y2751_class1.
png



8864_idx5_x150
1_y2801_class1.
png



8864_idx5_x155
1_y2401_class1.
png



8864_idx5_x155
1_y2451_class1.
png



8864_idx5_x155
1_y2501_class1.
png



8864_idx5_x155
1_y2551_class1.
png



8864_idx5_x155
1_y2601_class1.
png



8864_idx5_x155
1_y2651_class1.
png



8864_idx5_x155
1_y2701_class1.
png



8864_idx5_x155
1_y2751_class1.
png



8864_idx5_x155
1_y2801_class1.
png



8864_idx5_x160
1_y2351_class1.
png



8864_idx5_x160
1_y2401_class1.
png



8864_idx5_x160
1_y2451_class1.
png



8864_idx5_x160
1_y2501_class1.
png



8864_idx5_x160
1_y2551_class1.
png



8864_idx5_x160
1_y2601_class1.
png



8864_idx5_x160
1_y2651_class1.
png



8864_idx5_x160
1_y2701_class1.
png



8864_idx5_x160
1_y2751_class1.
png



8864_idx5_x160
1_y2801_class1.
png



8864_idx5_x165
1_y2251_class1.
png



8864_idx5_x165
1_y2301_class1.
png



8864_idx5_x165
1_y2351_class1.
png



8864_idx5_x165
1_y2401_class1.
png



8864_idx5_x165
1_y2451_class1.
png



8864_idx5_x165
1_y2501_class1.
png



8864_idx5_x165
1_y2551_class1.
png



8864_idx5_x165
1_y2601_class1.
png



8864_idx5_x165
1_y2651_class1.
png



8864_idx5_x165
1_y2701_class1.
png



8864_idx5_x165
1_y2751_class1.
png



8864_idx5_x170
1_y2251_class1.
png



8864_idx5_x170
1_y2301_class1.
png



8864_idx5_x170
1_y2351_class1.
png



8864_idx5_x170
1_y2401_class1.
png

Breast Histopathology Dataset

Breast Histopathology		本機磁碟	
..	檔案資料夾	2020/5/28 ...	
cnn.m	M 檔案	2,324	885
confusionmatStats.m	M 檔案	2,548	833
image2datastore.m	M 檔案	462	269

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

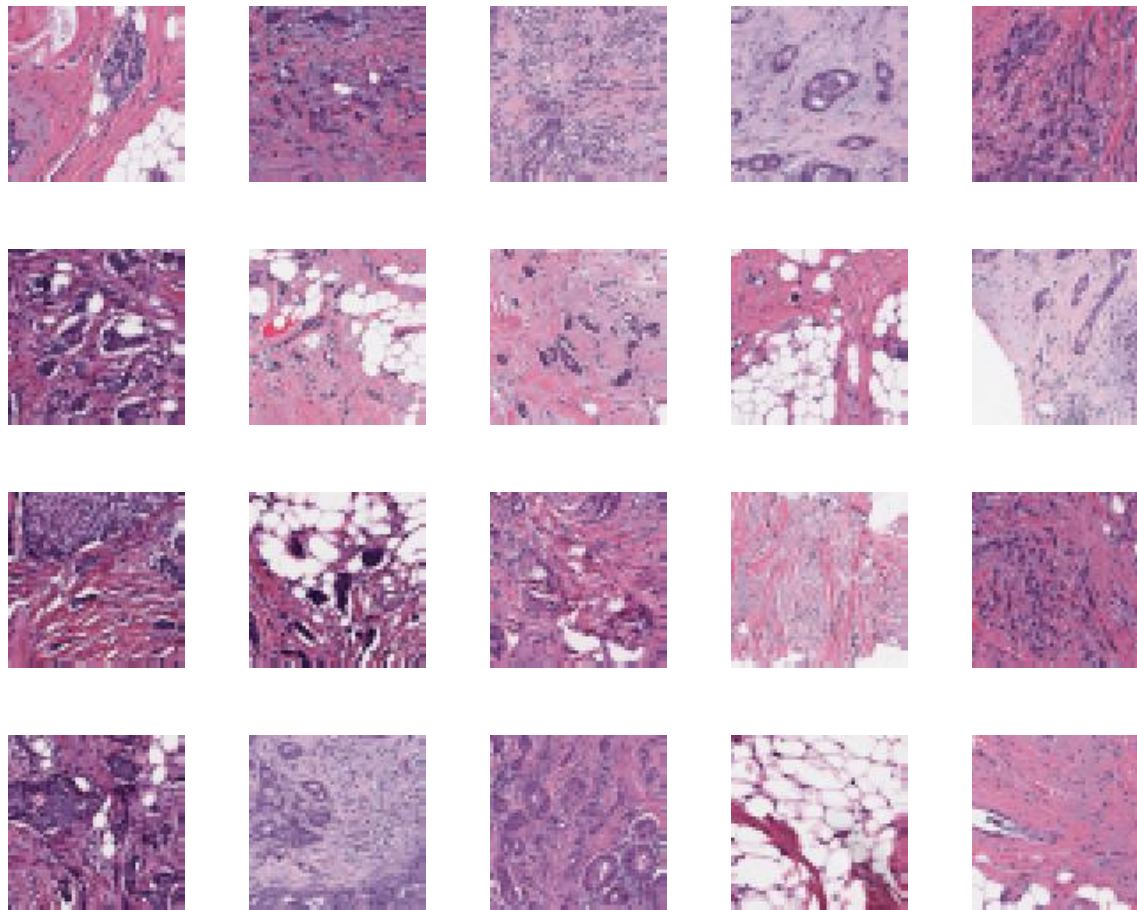
- `img =
imread('8864_idx5_x1_y2251_class0.png');`
- `imshow(img)`



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

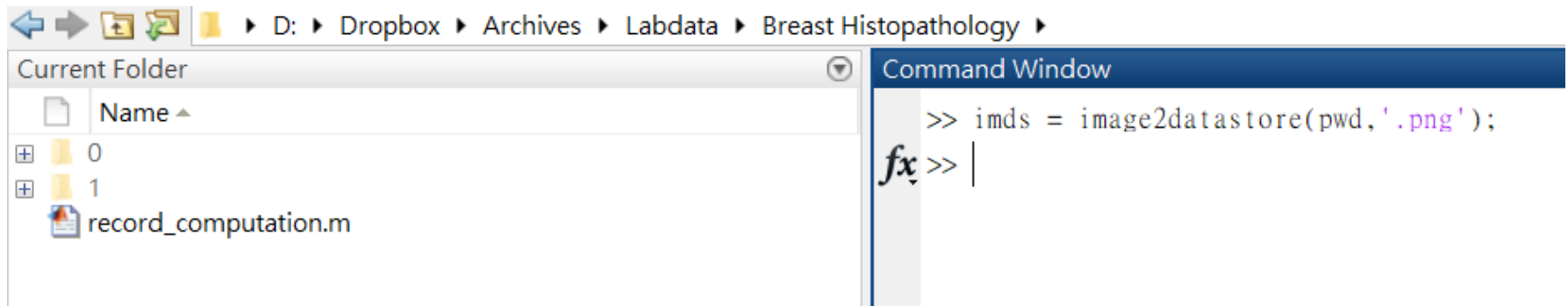
```
function imds = image2datastore(datapath,file_ext)

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

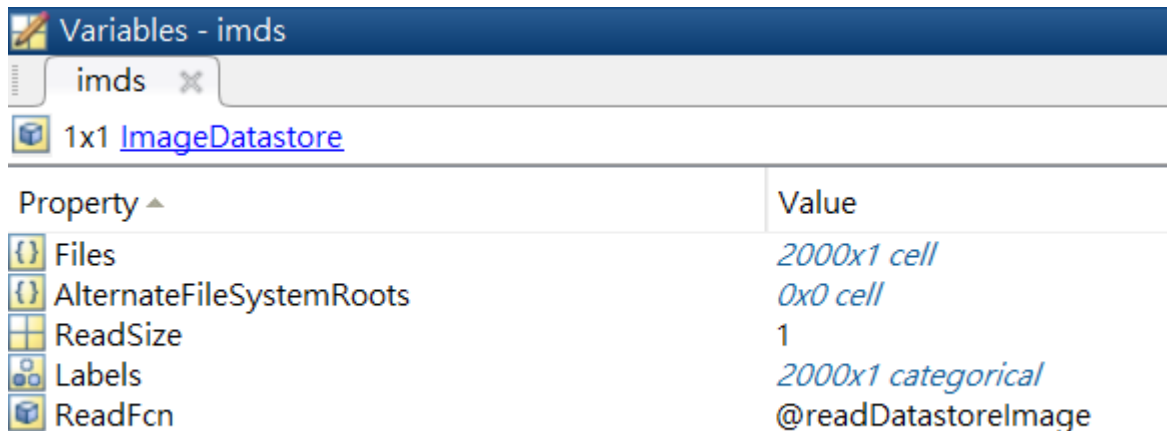
% Create image datastore using foldername and input file extension
filelocation = {};
for i=1:length(dinfo)
    filelocation{i} = [datapath '\' dinfo(i).name];
end
imds = imageDatastore(filelocation,'FileExtensions',file_ext,'LabelSource','foldernames');
end
```

Create Image Datastore

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



The screenshot shows the MATLAB interface. The top navigation bar indicates the current folder is `D:\Dropbox\Archives\Labdata\Breast Histopathology`. The file explorer on the left shows a folder named '0', a folder named '1', and a file named 'record_computation.m'. The Command Window on the right displays the command `>> imds = image2datastore(pwd, '.png');` and the prompt `fx>> |`.

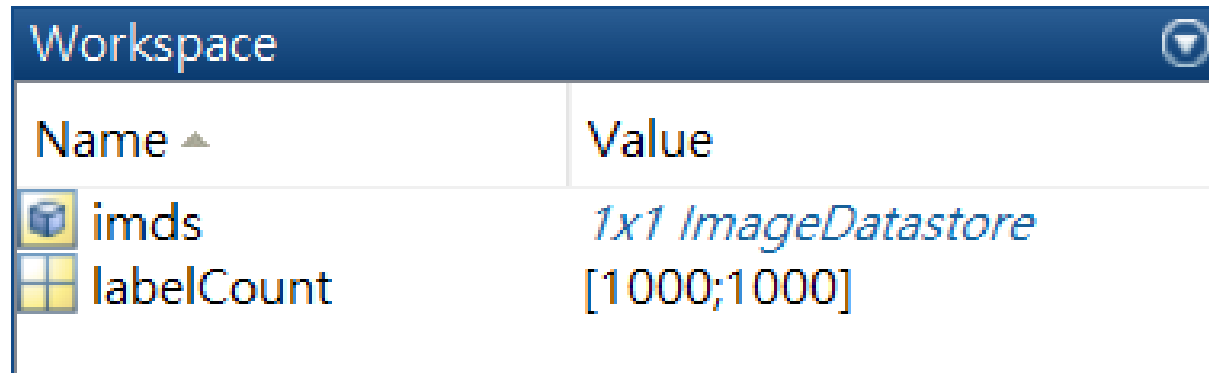


The screenshot shows the MATLAB Variables window for the variable 'imds'. It is a 1x1 ImageDatastore. The properties and their values are as follows:



Property	Value
Files	2000x1 cell
AlternateFileSystemRoots	0x0 cell
ReadSize	1
Labels	2000x1 categorical
ReadFcn	@readDatastoreImage

Count Number of Images for Each Label

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



The screenshot shows the MATLAB Workspace window with a dark blue header containing the text "Workspace" and a dropdown arrow icon. Below the header is a table with two columns: "Name" and "Value". The "Name" column has a small triangle icon next to it. There are two rows of data:

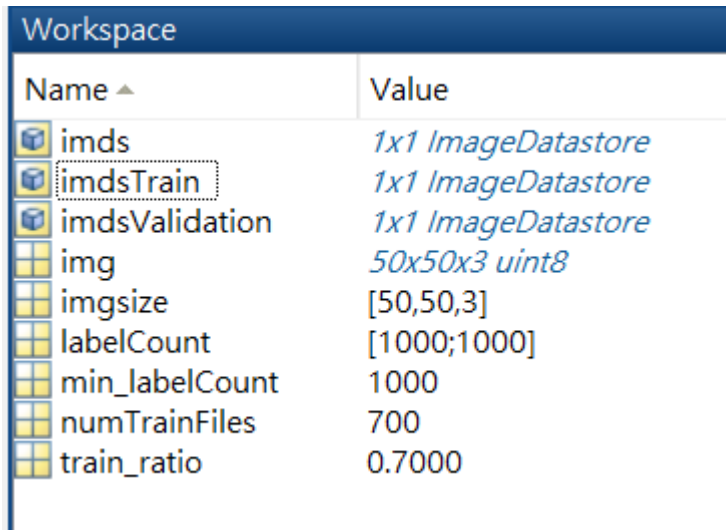
Name ▲	Value
 imds	<i>1x1 ImageDatastore</i>
 labelCount	[1000;1000]

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');`

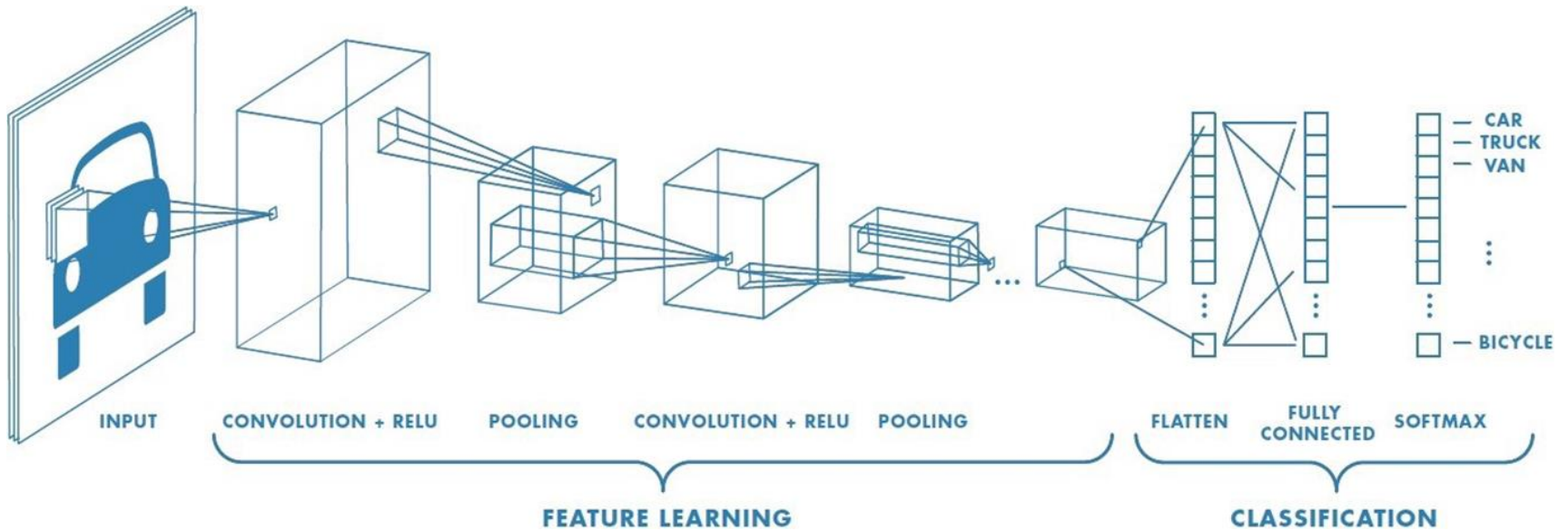


The screenshot shows the MATLAB Workspace window with a table of variables and their values. The 'imdsTrain' variable is highlighted with a dashed border.

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 Architecture

- 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];

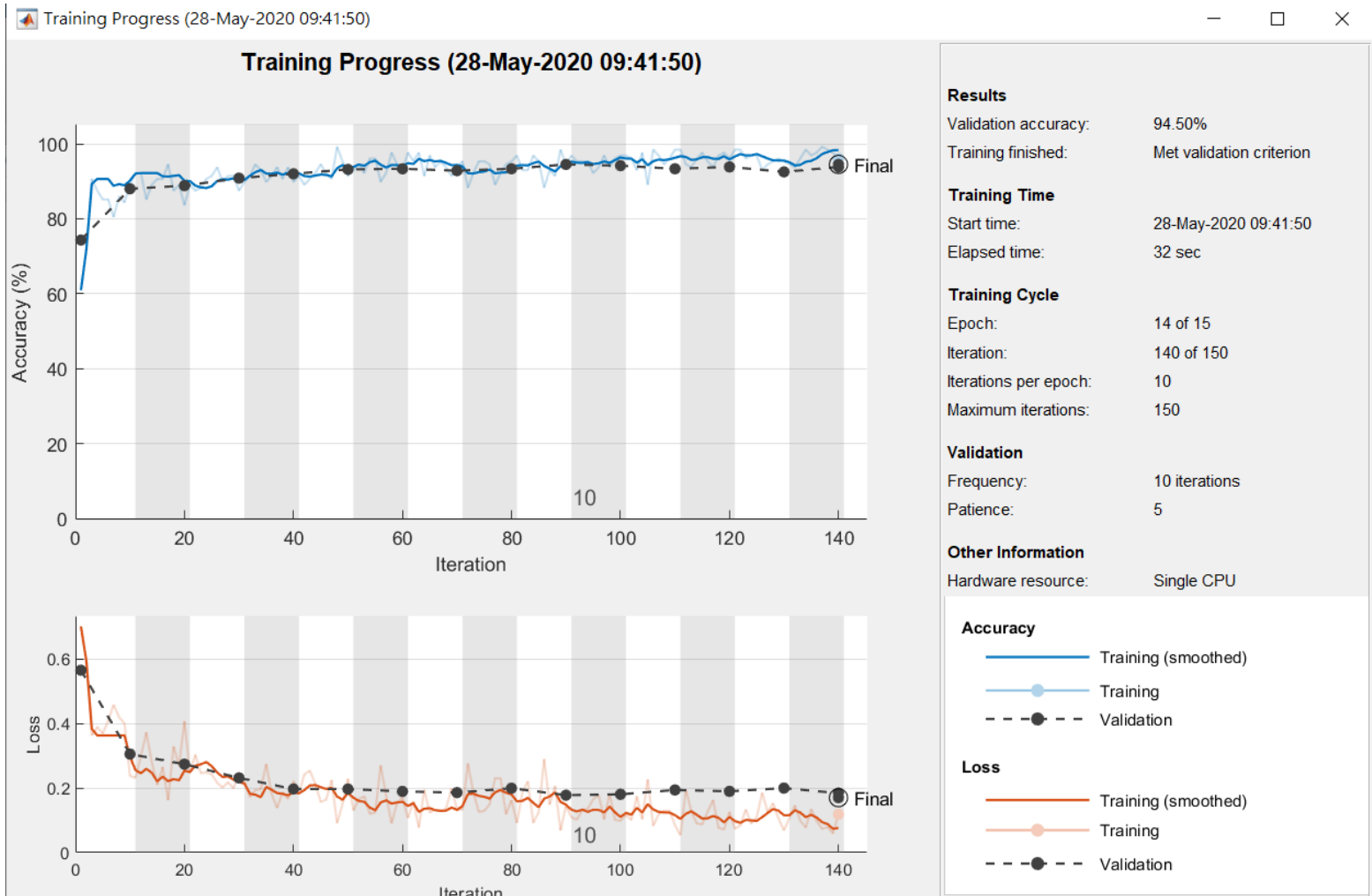
Specify Training Options

- `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

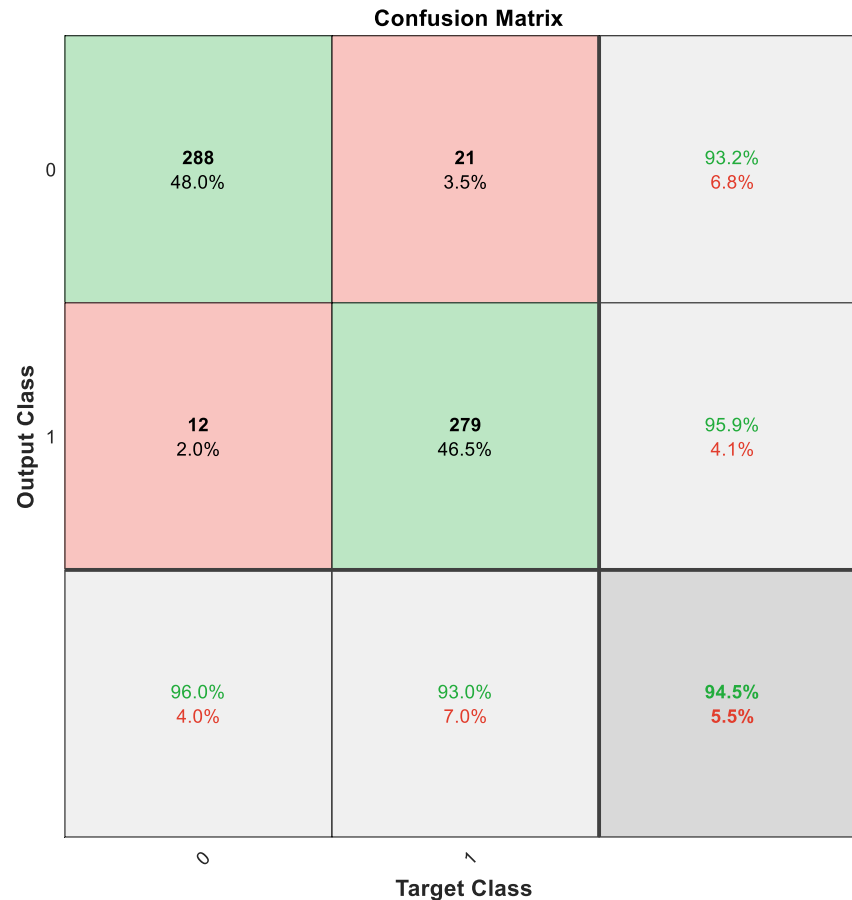


Compute Accuracy

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

Plot Confusion Matrix

- `plotconfusion(YValidation,YPred)`



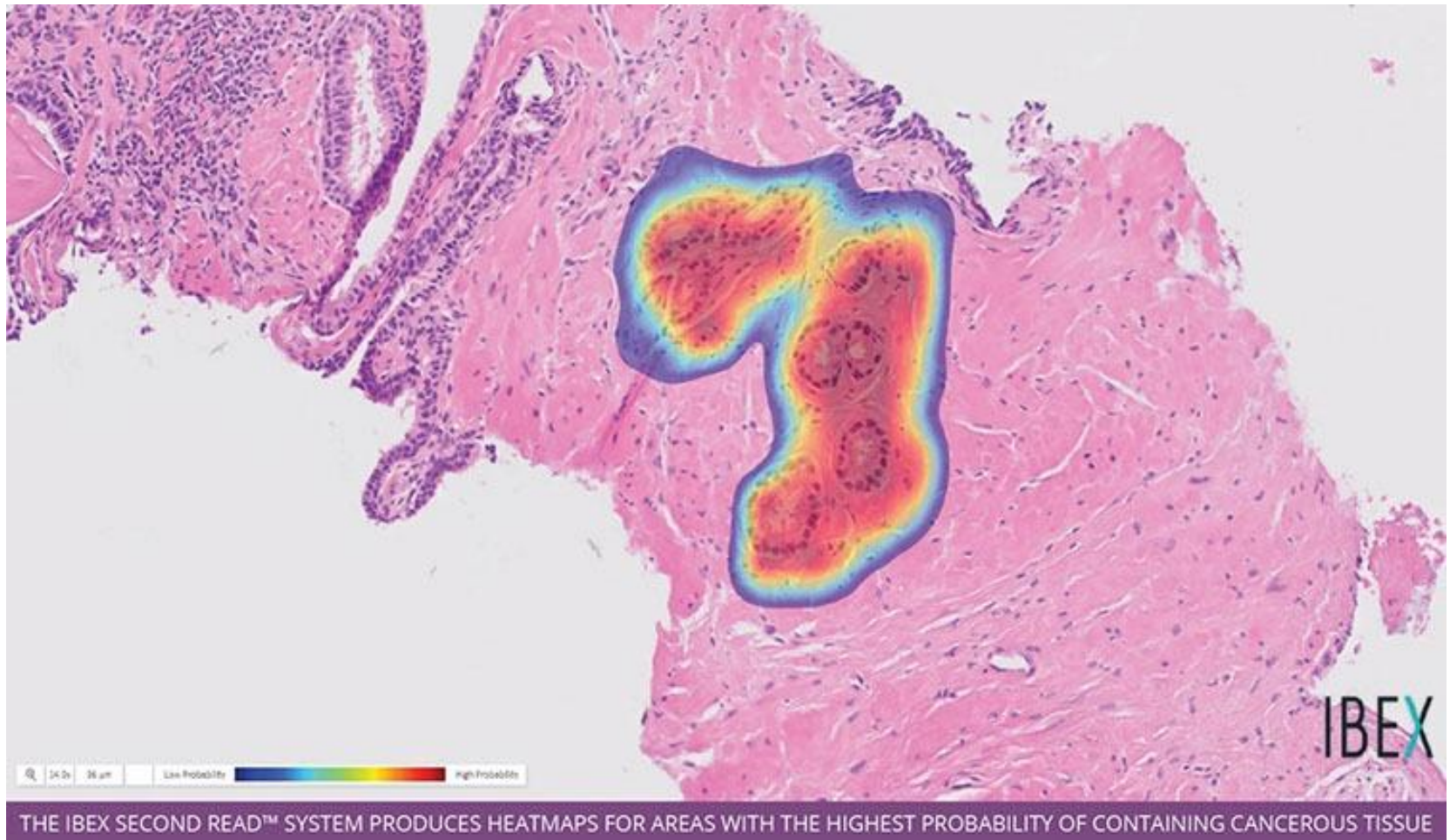
A CNN Code

```
cnn.m × record_computation.m × +
1 % cnn.m
2 % Dr. Albert C. Yang, MD, PhD. 2019/10/17
3 % Laboratory of Precision Psychiatry (http://www.precisionpsychiatry.org)
4
5 function [net netinfo nstats] = cnn(imds,train_ratio,show_progress,filter_size,num_filters)
6
7 if nargin<3 || isempty(show_progress)==1
8     show_progress = 1;
9     plots_option = 'training-progress';
10 end
11
12 if nargin<4 || isempty(filter_size)==1
13     filter_size = 4;
14     plots_option = 'training-progress';
15 end
16
17 if nargin<5 || isempty(num_filters)==1
18     num_filters = 8;
19     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

