FINAL Assignment
CS378/M392C/CSE392 (Fall 2018)
DUE Dec 10, 2018

For this take home final (project) you will exercise your research expertise in machine learning. **You are allowed to work in pairs. If you choose to do so, please declare your partners as you select your topic and outline your research goals. Your selection must be reported by email to me by Nov 21.**

Your final assignment todos are as follows. You must understand and summarize the algorithmic solution of the one of the linked papers. Additionally you should apply and present the results of the solution using the dataset we have provided (see below). You are additionally encouraged to research the problem solved by the suggested paper. Please report additional related papers you consulted and provides a “better” solution . A “better” solution is one that demonstrably yields computationally superior (accuracy vs speed) tradeoff solution for the same (similar) problem. The solution could be entirely “new”. Your welcome to come discuss it with me at anytime. On the last day of classes (Dec 10), you shall be asked to present details of your final solution, and submit a report, with an accuracy/speed tradeoff analysis. The format of the report must conform to the ICML paper submission format. For uniformity, please do not apply any other format. (You can download the Latex template the link here). Your main report should not exceed 8 pages. The bibliography should be listed at the end of the main report. The main report (should similar to other ICML papers) should include the following:

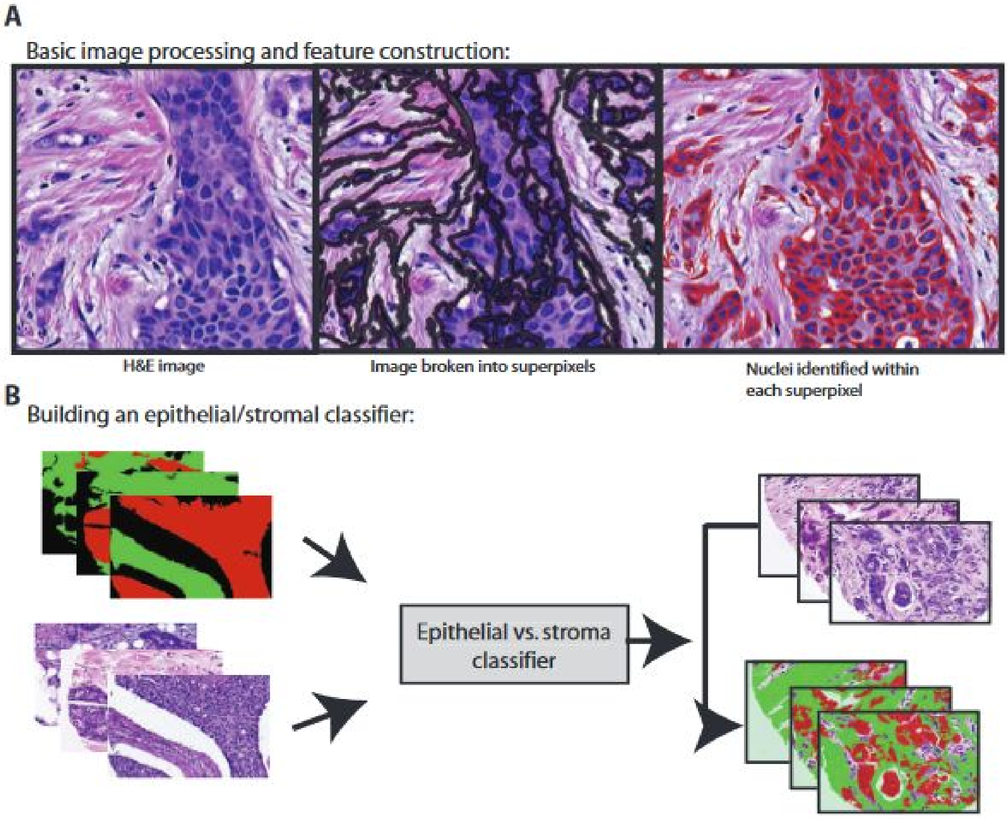
- Summarize the mathematical formulation and the algorithm presented to solve the specific problem, including theoretical accuracy/speed tradeoffs.
- Report, analyse and compare the results of your programming and verification. Use trade-off charts, plots, tables, figures etc.
- Provide a conclusion of what are the main bottlenecks to performance, and how you may attempt to resolve them, and thereby provide a better solution.

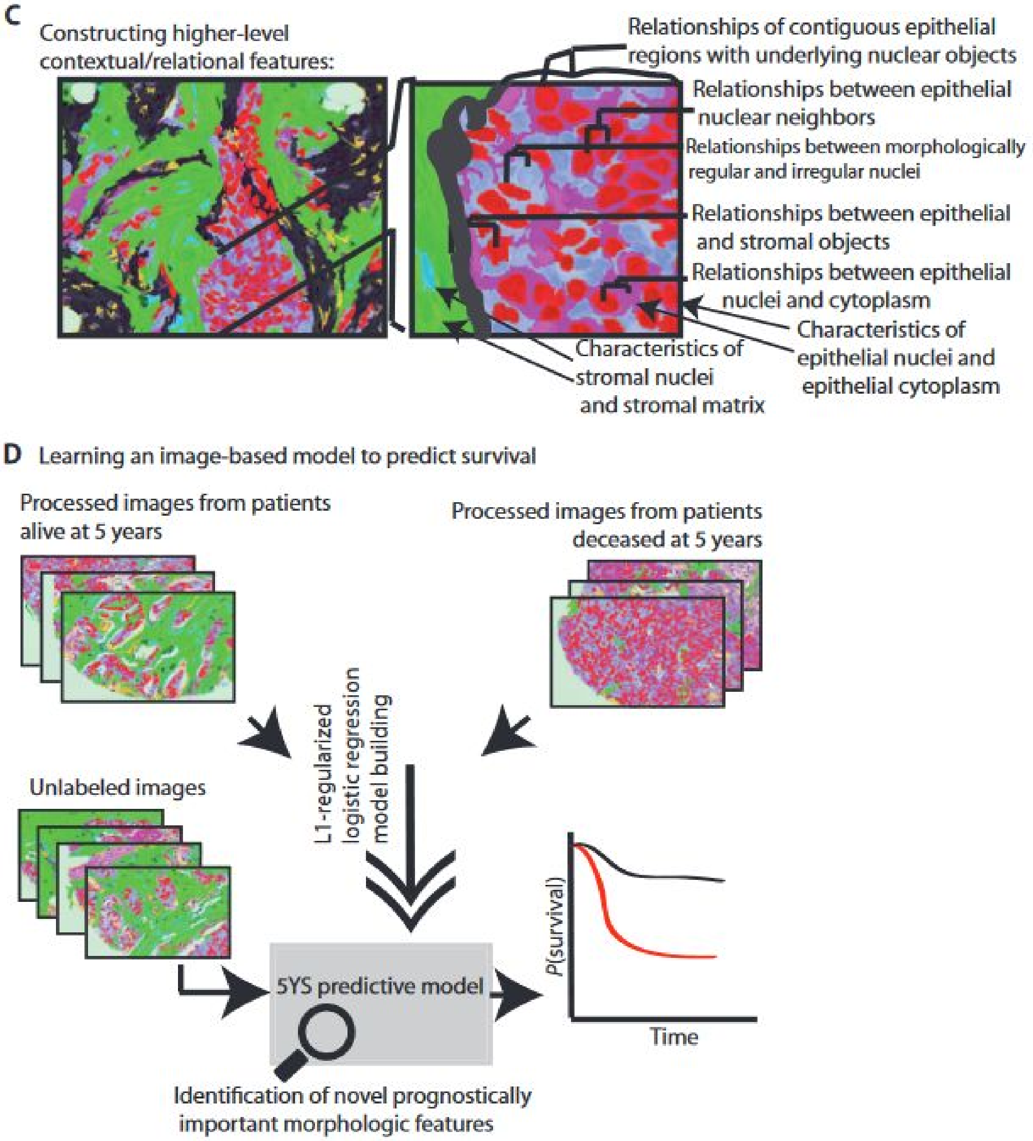
If your solution is deemed publishable, I shall be in touch after Dec 10. There is a deadline of Jan 23, 2019 for e.g. to submit to ICML 2019

Problem Domain and the Data Set

Please see [2,3] for background and to get conversant with the histopathology of stain image data, and your project’s problem domain. The descriptive figures A - D (from [3]) also explains their computational approach for “learning an image-model for pathology predictions”.

Disease diagnosis has been performed through the decades by a human pathologist who visually observes the stained specimen on a slide glass and using a microscope. In recent years, attempts have been made to capture the entire tissue slide with an optical scanner and save it as a digital image. The data collection you will work with is such a set of digitally scanned stain images The primary problem of the digital pathologist is to classify the different types of tissue samples to reveal different pathologies or disease states. The challenge is also of big data, as a single high resolution stain image can be 2K x 2K RGB pixels, and in some cases of size 20K x 20K RGB pixels. There are also related hyperspectral datasets called FTIR which are 1K x 1K images with each pixel being of 1500 spectral channels.

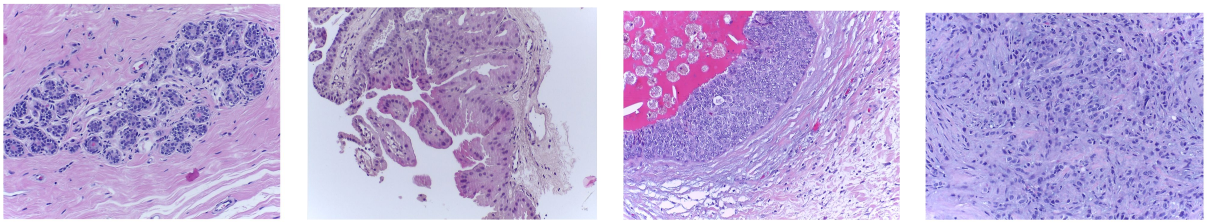

A
Basic image processing and feature construction:
H&E image
Image broken into superpixels
Nuclei identified within each superpixel
B
Building an epithelial/stromal classifier:
Epithelial vs. stroma classifier


C
Constructing higher-level contextual/relational features:
Relationships of contiguous epithelial regions with underlying nuclear objects
Relationships between epithelial nuclear neighbors
Relationships between morphologically regular and irregular nuclei
Relationships between epithelial and stromal objects
Relationships between epithelial nuclei and cytoplasm
Characteristics of epithelial nuclei and epithelial cytoplasm
Characteristics of stromal nuclei and stromal matrix
D Learning an image-based model to predict survival
Processed images from patients alive at 5 years
Processed images from patients deceased at 5 years
L1-regularized logistic regression model building
Unlabeled images
5YS predictive model
P(survival)
Time
Identification of novel prognostically important morphologic features

Data Sets [1]:

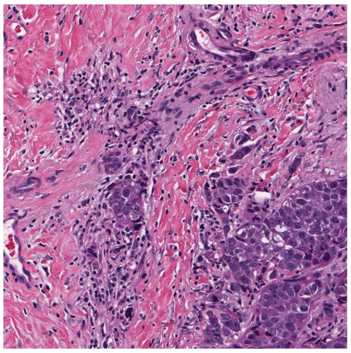
Bioimaging Challenge 2015 Breast Histology Dataset:

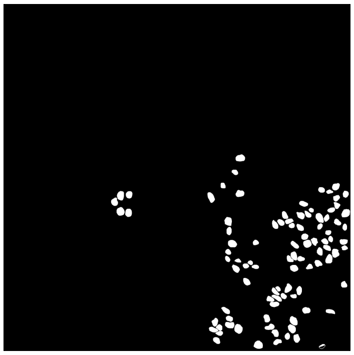
The datasets contains around 250 high resolution image as training set, and 35 as the test set, we will only give you the training test labelings, that is one of the four tags as indicated below:



Left to right: Normal initial, Benign Initial, In situ Initial, Invasive Initial

Nuclei Segmentation Dataset:
Detecting nuclei within H&E stained estrogen receptor positive (ER+) breast cancer images The dataset consist of 143 images ER+ BCa images scanned at 40x. Each image is 2,000 x 2,000. Across these images there are about 12,000 nuclei manually segmented. The format of the files is:
12750_500_f00003_original.tif: original H&E image



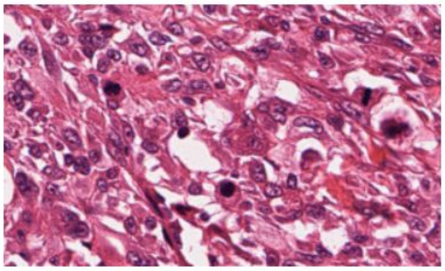


12750_500_f00003_mask.png: mask of the same size, where white pixels are nuclei

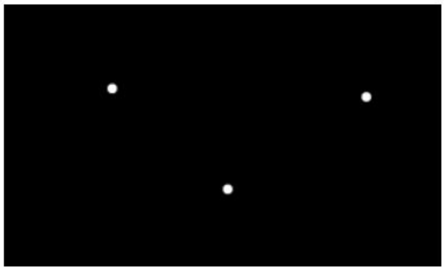
Mitosis Detection Dataset:
The dataset consist of 311 images of size 2,000 x 2,000 @40x selected from 12 breast cancer (BCa) patients. In total there are 550 mitosic centers expertly identified using a focal microscope.

The format of the files is:01_01.tif: original H&E image
01_01.csv: x,y coordinates of where the mitotic center are

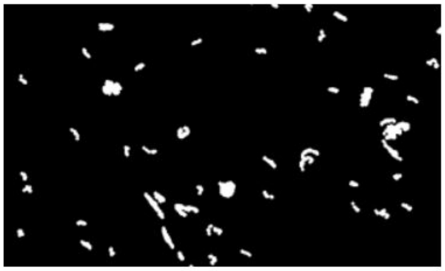
<u>01_01_pc.png:</u> a helper image which places circles at mitotic centers
<u>01_01_cmask.png:</u> the blue ratio segmentation mask



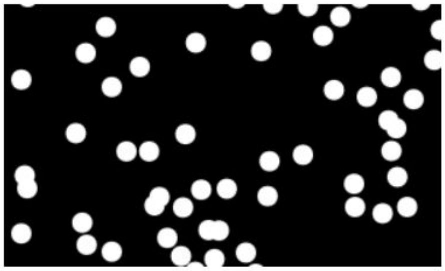
Original Image



Centers Marked of Mitosis



Blue Ratio segmented Image

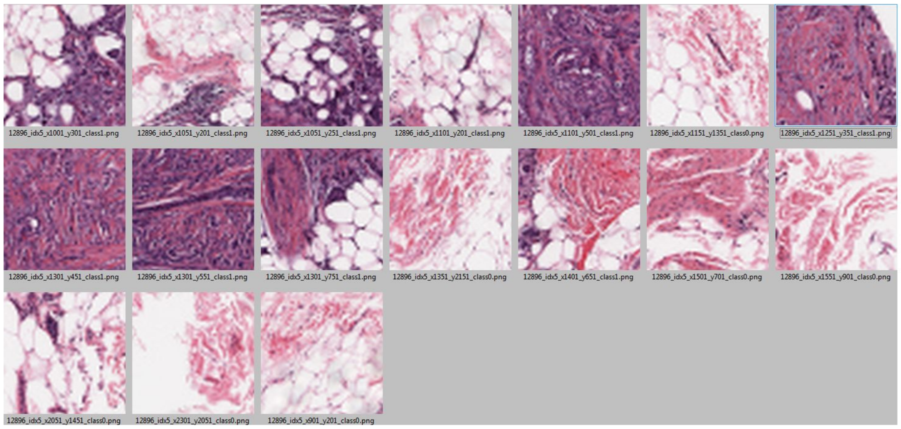


Dilated Version of Blue Ratio

Invasive Ductal Carcinoma (IDC) Identification Dataset:
This is a binary classification problem 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:

u_xX_yY_classC.png   — > example 10253_idx5_x1351_y1101_class0.png

Where **u** is the patient ID (10253_idx5), **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 IDC.



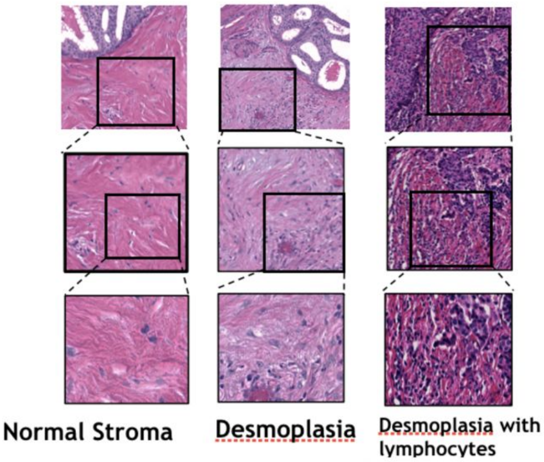
12896_idx5_x1001_y301_class1.png
12896_idx5_x1051_y201_class1.png
12896_idx5_x1051_y251_class1.png
12896_idx5_x1101_y201_class1.png
12896_idx5_x1101_y501_class1.png
12896_idx5_x1151_y1351_class0.png
12896_idx5_x1251_y351_class1.png
12896_idx5_x1301_y451_class1.png
12896_idx5_x1301_y551_class1.png
12896_idx5_x1301_y751_class1.png
12896_idx5_x1351_y2151_class0.png
12896_idx5_x1401_y651_class1.png
12896_idx5_x1501_y701_class0.png
12896_idx5_x1551_y901_class0.png
12896_idx5_x2051_y1451_class0.png
12896_idx5_x2301_y2051_class0.png
12896_idx5_x901_y201_class0.png

Project Problems :

There are two overlapping problem scenarios stated below. You can pick either one or also combine significant flavours of (1) and (2). For either problem scenario, I suggest you additionally research the the survey [4] to apprise you of the range of prior computational work. I have additionally designated a primary dataset to use to demonstrate your solution. This however in no way precludes your use of other datasets to train and or to demonstrate your solution. If you have difficulty visit dataset source, you can download the copy at here.

(1) Multi-scale Classification and Grading of Pathological Tissue Stains

Develop a data driven learned tissue stain classifier and tissue grader. Your goals are to go beyond what you did in your past assignment projects using Kernel-PCA. Namely, the advantage of your new learned classifier is to find the optimal balance of the sparsest dictionary of features and the maximal number of correct tissue grades (classes) that you would correctly discriminate amongst. You should examine the tissue feature sets and the relationships between them, at multiple macro to sub-micron scales. Sub-pixel data unmixing could also be explored. Deep Learning methods can also be utilized (see [5] for e.g).



An appropriate dataset for learning and testing is available from **Bioimaging Challenge 2015 Breast Histology Dataset**. You should additionally apply your algorithm to the **Invasive Ductal Carcinoma (IDC) identification Dataset.**

(2) Compressive Sensing for Robust Sparse Recovery of Image Statistics

To provide a solution to the big data problem, another goal is to compressively sense/query the stain data images for sparse approximate recovery of   sparse features, their properties, and statistical relationships. For instance, the arrangement of nuclei might be an indicator of the pathology status of the stain image, but the nuclei only fill up a small portion of the image. In this task, you are required to design a compressive sensing algorithm that best recovers the nuclei mask positions. You could also combine this compressive technique with the M-estimation techniques, similar to the Alternating Maximization with Latent Variable Model (AM-LVM) probabilistic learning method [7]. See also the survey article on compressive sensing [6].

You should apply your algorithm to the **Nuclei Segmentation Dataset.** Of course please also consider the **Mitosis Detection Dataset.**

References

1. DataSet Source : http://www.andrewjanowczyk.com/deep-learning/ Bioimaging Challenge 2015 Breast Histology Dataset : https://rdm.inesctec.pt/dataset/nis-2017-003
2. Dana Carmen Zaha, “Significance of immunohistochemistry in breast cancer”, *World J Clin Oncol* 2014 August 10; 5(3): 382-392
3. Andrew H. Beck, Ankur R. Sangoi, Samuel Leung, Robert J. Marinelli, Torsten O. Nielsen, Marc J. van de Vijver, Robert B. West, Matt van de Rijn, Daphne Koller, “Systematic Analysis of Breast Cancer Morphology Uncovers Stromal Features Associated with Survival” www.ScienceTranslationalMedicine.org 9 , 2011, Vol 3, Issue 108
4. Komura, Daisuke, and Shumpei Ishikawa. "Machine learning methods for histopathological image analysis." *Computational and Structural Biotechnology Journal* 16 (2018): 34-42.
5. Teresa Araujo, Guilherme Aresta, Eduardo Castro, Jose ́ Rouco, Paulo Aguiar, Catarina Eloy, Antonio Polonia, Aurelio Campilho “Classification of breast cancer histology images using Convolutional Neural Networks”, PLOS One, https://doi.org/10.1371/journal.pone.0177544
6. Baraniuk, Richard G. "Compressive sensing [lecture notes]." *IEEE signal processing magazine* 24.4 (2007): 118-121.
7. Sivaraman Balakrishnan, Martin J. Wainwright, and Bin Yu. Statistical Guarantees for the EM Algorithm: From Population to Sample-based Analysis. *Annals of Statistics*, 45(1):77–120, 2017

Hyperspectral Image Data: BR-1003 FTIR dataset:

This is, rigorously speaking, not a dataset, it contains a snapshot of 100 tissue samples in one hyperspectral FTIR data. The size of the data is 11620 by 11620 by 1506. The number 1506 represents the band number of this data. For example, RGB image will have a band number 3 while a grayscale image will only have one band.

Researchers also provide stain image (which you have seen in the previous dataset). Simply speaking, stain image is the dyed result for tissue samples. The size of stain image is 42464 by 44107 RGB image each. The tissue sample contains classification result(source), row 1 and 2 belong to Hyperplasia, row 3 and 4 are atypical, row 5 to 8 belong to the malignant class and the last two rows are normal.

In addition to the original data, we will also provide you additional manual labeling mask for different types of tissues in this hyperspectral data. For example, we have the location in FTIR which represents Dense stroma. (Other type of labels are: Loose Stroma, Malignant, Normal, Others, and Reactive Stroma (Desmoplasia)).