Computer Assisted Classification of Adenocarcinoma

Research Summary

Student: Ofer Givoli
Instructor: Dr. Erez Berkovich

Contents

1 Introduction

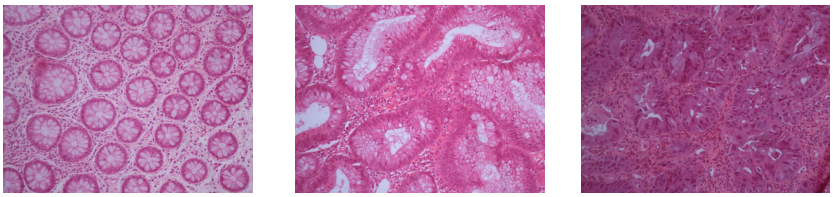
We have a dataset containing images that are the digital scanning of H&E stained tissues taken from the colon. Our dataset contains the following classified images:

1. 125 images of normal tissue.
2. 380 images of adenoma (almost cancerous).
3. 258 images of adenocarcinoma (cancerous).

The images are in 200x magnification, and their size is 1280x960 pixels. Some color related conditions may vary.

The project's goal was developing an algorithm that classifies such images into the above classes (normal tissue, adenoma, adenocarcinoma). We've decided on a set of features to extract from each image in order to train and use a classifier (using the machine learning approach).

The following figure demonstrates the images we work with:



The many circular objects that can be seen in the normal tissue image are called crypts. The crypts usually loose their round shape in the adenoma stage, and are usually completely destroyed in the adenocarcinoma stage.

As of this day, the detection of adenocarcinoma in H&E stained tissues is done manually by human pathologists. The time pathologists spend examining tissues from each patient is limited and expensive. Developing an automatic classifier as described above might be very beneficial for humanity.

A lot of research has been done in this field in recent years. Many researchers published the performance of classifiers they developed using a big variety of features, such as:

- *Statistical Features*[1]: Average, median, standard deviation, difference, Sobel and Kirsch filters, derivatives in the horizontal, vertical, and diagonal directions.
- *Co-occurrence Features*[1]: Angular second moment, contrast, correlation, variance, entropy, inverse difference moment, sum average, variance, and entropy, difference variance, and difference entropy
- *Object-Based (Structural) Features*: Mainly crypts related [5],[2],[7].
- *Other Features*: Operators motivated by human vision [6], Multiwavelets [3], Gabor based features [1].

The next section describes the features we chose to experiment with in this project.

2 Features Used

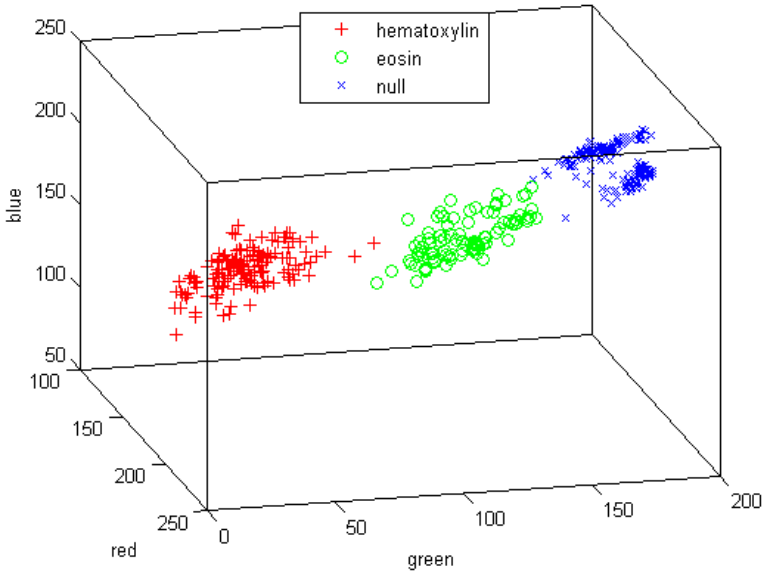
2.1 HematoxylinEosin Image (HE Image)

Many papers dealing with computer aided adenocarcinoma detection in H&E images were published. In almost all papers I've read (such as [4], [6], [3]) the authors worked with plain gray scale image. Authors took the original RGB images and simply converted them to gray images by taking the mean among the three channels. Our main goal in this project became coming up with a better approach.

Let us reflect on what is the best way to represent the information in an H&E image. Suppose we have an RGB image of an H&E stained tissue. The information that each pixel hold is how much did each of the two dyes (hematoxylin

and eosin) stain the area in the tissue that the pixel represent[1]. So suppose we could take the RGB image and convert it to a 2-channels image called HematoxylinEosin Image (HE Image). Each pixel in the HE image will hold a pair of values in the range $[0, 1]$ that would represent how much each of the two dyes stained the area of that pixel. We argue that this representation is superior to that of a simple gray image because it better represents the information that the original image holds.

In an RGB image, regions stained by hematoxylin tend to be bluish, and regions stained by eosin tend to be pink. The following figure show the RGB values of pixels that were manually identified as being stained by either hematoxylin, eosin, or none of them (we refer to such pixels as belonging to the class of the "null dye").



We believe HematoxylinEosin images provide a better representation of the available information than gray, RGB or even HSI images (as was done for example in [1]). Also, it might hold information that is completely lost when converting the RGB image to a gray image. Our hope is that HematoxylinEosin images will prove to be useful for many researchers dealing with H&E images, for extracting features that are object-based and otherwise.

Converting RGB to HE images

So given an RGB image I, how do we convert it to an HE image?

We've sampled many pixels in the original RGB images that we could manually determine to belong strictly to one dye: either hematoxylin, eosin, or the "null dye" (those pixels are those used in the previous figure). From now

[1] The staining of hematoxylin and eosin depends on pH level

on we'll refer to three dye classes: hematoxylin, eosin and null. For each pair of dye classes (of the 6 possible pairs), we've used Fisher's linear discriminant (FLD) to find a vector in RGB space that separates the two classes well. For each such vector, we also calculated the mean projection of each of the two dye classes on the vector. So given a new pixel p, we can find the projection of p on such a vector, and determine to which of the two dye classes p is closer to.

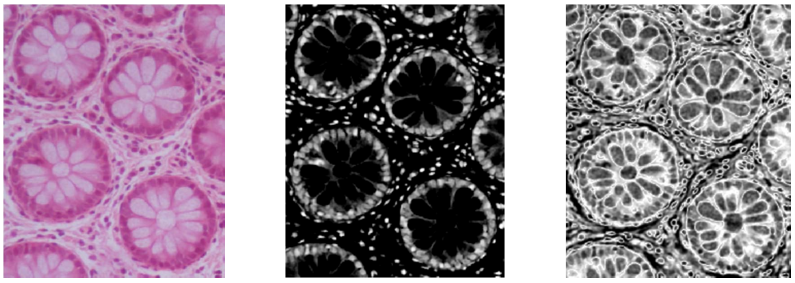
Since FLD works with two classes, and we have three, we did the following trick. Given a new pixel p, we checked each of the 6 dye class pairs, and for each pair we've declared a "winning" class: the class that p was closer to. Then we kept the two classes that won the most, and determined the HE pixel value according to the FLD vector between those two classes (ignoring the third "loosing class").

If the two winning classes are hematoxylin and eosin, then the two channels of the HE pixel are linearly determined according to how close the projection of the pixel was to the mean of each of the two classes on the FLD vector. The sum of the two channels is 1.

If one of the winning classes is the null dye, then the channel of the other winning class (either hematoxylin or eosin) in the HE pixel is determined as before, and the channel of the loosing class is set to 0.

Note: In rare cases there is no "loosing class" because every class won once and lost once. We're treating this situation as if the winning classes are hematoxylin and eosin.

The following figure demonstrate an HE image converted from an RGB image. Hematoxylin Image is the first channel, and Eosin Image is the second channel of the HE image.



Original (RGB) Hematoxylin Image Eosin Image

Note: The pixel data used for the FLD analysis can be regenerated when the lab equipment change, in order to tackle the variance during the image acquisition process.

2.2 Features Extracted

We extracted from the images the following features.

1. We have extracted 64 Gabor based numeric features using GIST[2]. GIST is a Matlab library that extracts a vector of features from an input image, that are based on Gabor filters and are taken from multiple scales and orientations. We've extracted the GIST features from:

   (a) the gray images.

   (b) each of the two channels of HE images.

2. We have developed an object-based feature called "Circles_1". This numeric feature is meant to count the number of normal crypts in the image. To extract this feature we basically count the objects in the image that are similar to a circle (and larger than some threshold). The circular objects counted are meant to represent the crypts. This is useful because in images of normal tissue there are a lot of crypts shaped almost as a perfect circle. In images of cancerous tissues (adenoma, adenocarcinoma) circular crypts are rare. We computed this feature using existing code from a demo in Matlab called "Identifying Round Objects". On a given RGB image I the feature is computed by the following algorithm:

   (a) Convert I to a gray image (by simply taking the mean of the three channels).

   (b) Convert I to a binary image using a threshold determined by Otsu's method.

   (c) $I \leftarrow \bar{I}$ (meaning: flip all the pixels of I)

   (d) Remove from I white elements that are less than 1000 pixels in size.

   (e) Fill up holes in I, where holes are defined as black areas that are not connected to the image borders by a continuous path of black pixels.

   (f) Perform on I the morphological operator 'open' where the structuring element is a disk with a radius of 50 pixels.

   (g) At this point I is a binary image in which (hopefully) the crypts are white and most of the rest of the image is black. So we treat continues white regions in the image as elements, and we wish to detect the round elements (which hopefully are crypts). So for each element e in the image, calculate the "circle similarity" ratio:

   $$s = \frac{4\pi \cdot \text{area}\,(e)}{(\text{perimeter}\,(e))^2}$$

   For circles we'll get $s = 1$, and for elements with shapes that are not a circle we'll get $s < 1$. Intuitively, as the shape is less concentrated around its center of mass, s will be larger. The more the shape is similar to a circle, the higher s will be. Using the threshold 0.85 we determine which elements in the image are similar enough to a circle. We count those elements and that is the feature value returned.

[2] http://people.csail.mit.edu/torralba/code/spatialenvelope/

3. We have developed an object-based set of features called "CryptsFeatures_1". We take advantage of the HE representation and try to detect the crypts more accurately and more robustly than was done in "Circles_1" (that only used the gray images) . Then we extract the following features:

   (a) The number of crypts detected.

   (b) Median "circle similarity" (see definition above) of the crypts.

   (c) Median area (in pixels) of the crypts.

   Given an HE image I, the detection of the crypts is done by the following algorithm:

   (a) Calculate a binary image cryptsOuterLayer that represents all the pixels in I where the hematoxylin channel is above some threshold. The white pixels in cryptsOuterLayer are meant to represent the outer layer of the crypts (not accurate).

   (b) Calculate a binary image notUniform that represents all the regions in I that were not uniformly stained by one of the two dyes. This is done by performing a series of morphological operators (basically we do 'close' operation on the negation of each of the two channels, and then do logical 'AND' operation on the two images received).

   (c) Calculate a binary image 'crypts' to be: $not(or(notUniform, cryptsOuterLayer))$. The image 'crypts' now roughly represents the crypts in I. We perform on crypts some additional morphological operators:

      i. throw away elements with less than 1000 pixels.

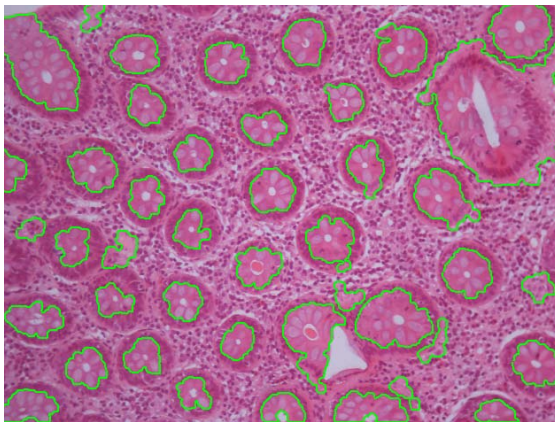
      ii. fill holes (as defined above).

      iii. use the 'open' operator where the structuring element is a disk with a radius of 10.

      iv. throw away elements with less than 2000 pixels.

      Finally, we arrive at a binary image that for some images represent the crypts very accurately.

   The following figure demonstrate the crypts detection done by the algorithm described above (detected regions are marked by green borderlines) :



Note that there are images (especially images of adenoma and adenocarcinoma) for which the algorithm works very poorly.

4. We have developed an object-based feature called "HematoxylinCircles_1", which is similar to "Circles_1", but uses the hematoxylin channel of the HE image instead of the gray image. I haven't finished working on it (additional parameter tuning should be done).

The following figure describes the performance achieved with the AdaBoost+C4.5 classifier using the individual features and sets of features:

Feature Name	AdaBoost-J48; Leave one out			AdaBoost-J48; 2:1 train-test sets		
	Area under ROC (NormalTissue)	Sensitivity (NormalTissue)	Specificity (NormalTissue)	Area under ROC (NT)	Sensitivity (NT)	Specificity (NT)
Circles_1	0.898	78.8%	98.4%			
GIST_1	0.871	49.4%	93.5%			
HematoxylinCircles_1	0.735	49.4%	97.9%	0.859	56.7%	100.0%
GIST_with_FLD_HematoxylinImage	0.947	75.3%	96.3%	0.946	80.0%	94.6%
GIST_with_FLD_EosinImage	0.786	43.5%	95.4%	0.698	20.0%	89.2%
GIST_with_FLD_HematoxylinAndEosinImage	0.960	70.6%	96.1%	0.922	70.0%	96.4%
CryptsFeatures_1	0.823	52.9%	95.6%	0.823	50.0%	89.2%
Set of Features						
Circles_1 + GIST_with_FLD_HematoxylinAndEosinImage	0.995	87.1%	98.8%	0.996	93.3%	99.4%
FeatureSet_001: Circles_1 + HematoxylinCircles_1 + GIST_with_FLD_HematoxylinAndEosinImage	0.995	91.8%	98.6%	0.975	66.7%	99.4%
FeatureSet_002: Circles_1 + HematoxylinCircles_1 + GIST_with_FLD_HematoxylinAndEosinImage + CryptsFeatures_1	0.987	92.9%	99.1%	0.993	83.3%	97.4%

An interesting result is that extracting Gabor features (using GIST) from both channels of the HE images yields better results than extracting the Gabor features from the gray images.

3 Classifier Selection

We have used Weka (a machine learning java library) for compering the performance of many known classifiers from the field of machine learning. The

following figure describe the performance of the classifiers compared, using the features:

1. Circles_1
2. GIST_with_FLD_HematoxylinAndEosinImage

The results (area under ROC, sensitivity, specificity) are defined by the Normal Tissue concept. The experiment was done using leave-one-out.

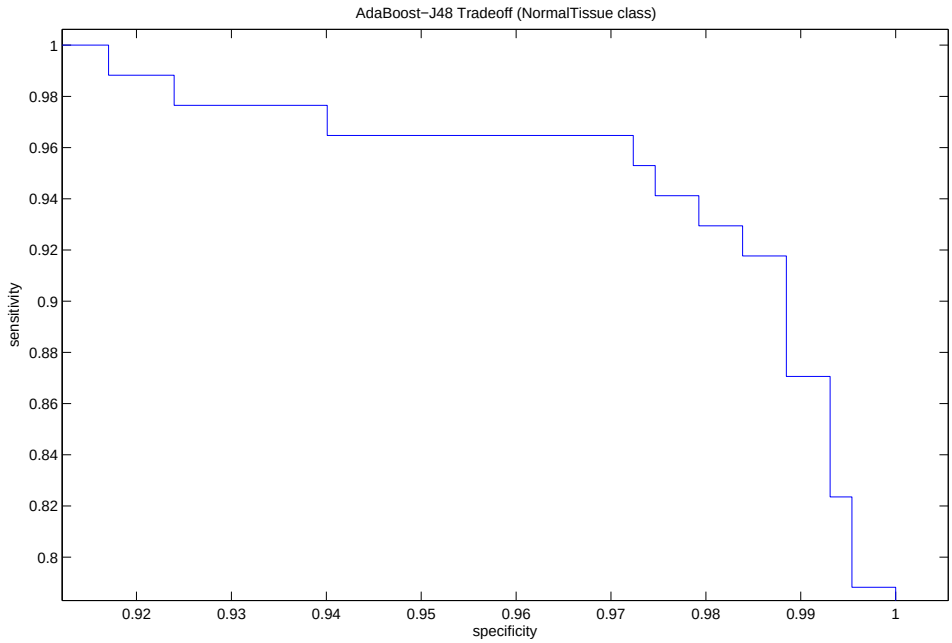
Classifier (Leave one out)	Area under ROC	Sensitivity	Specificity
AdaBoost-J48	0.995	87.1%	98.8%
AdaBoost-Naive Bayes	0.868	91.8%	79.5%
AdaBoost-KNN, K=3	0.934	87.1%	95.6%
AdaBoost-SVM, gaussian kernel	0.878	75.3%	99.1%
KNN, K=3	0.951	88.2%	95.4%
Naive Bayes	0.931	91.8%	79.5%
KNN, K=1	0.910	84.7%	97.2%
J48	0.880	78.8%	95.9%
SVM, linear kernel	0.877	76.5%	98.8%
SVM, gaussian kernel	0.874	75.3%	99.5%

The results show that AdaBoost using J48 (which is an implementation of the well known C4.5 decision tree classifier) had the best performance. So we chose that classifier to work with.

4 Classifier Performance Evaluation

The following ROC curve was generated using the AdaBoost+C4.5 classifier with the features:

1. Circles_1
2. GIST_with_FLD_HematoxylinAndEosinImage



5 Conclusion

We've build a classifier that recognizes cancerous tissues (adenoma + adenocarcinoma) that is capable of achieving sensitivity of 98.8% and specificity of 87.1%.

The novelty of our work is mainly presenting the HE image representation. As far as we know, this has never been done before. Extracting Gabor features from both channels of the HE images yielded better results than extracting the Gabor features from the gray images. We hope the HE image representation will prove useful for many researchers dealing with images of tissues stained by H&E.

An additional result that might be useful to the community is that among all classifiers we've tried, AdaBoost with C4.5 yielded the best performance (for the features we've used).

6 Ideas for Extending the Research

6.1 Extending our Research to Larger Images

We can ask Dr. Sabo to use the dotSlide digital microscope to acquire larger images (possibly covering the entire tissue), with the same resolution as our current images. Then we could do the following:

1. We will take large images that we know to contain both normal tissue and cancerous areas. We'll split each large image into many (~1000 ?) smaller images (same size as the images used in our basic research). Then we'll simply use our classifier from the basic part, to classify the smaller images. The results will define a partitioning of the large images into regions with different cancerous state. This could have useful practical uses.

2. We could extend the set of features that we examine, and consider features of lower scale images (considering the architecture of the tissue). We could perhaps use the features from the smaller images to define "super-features":

   - spatial features in a large image based on features of the smaller sub-images (of the large image).

   From the papers I've read it seems that the territory of using features from a large range of scales (in histopathological images) is unexplored. Everything we'll do might be novel.

6.2 Automatically Searching the Domain of Features Subsets

We can try a feature selection approach. Suppose we have a huge set of features F. The set F can even be infinite. For example, suppose each feature from our basic research can be used as a family of features (all in F), by letting some parameters vary in some ranges, etc.. Now suppose we develop an algorithm whose goal is to find a "good" subset of features $S \subset F$, such that when we take a learner and use it with the features in S, we get an accurate classifier.

We can take this idea one step further. When using object-based features, we can add another level of flexibility to our algorithm (formally, adding a huge amount of possible features to F). We will let our algorithm search for a "good" object-based representation, such that when using that representation, "useful" features can be extracted.

References

[1] S. Doyle, C. Rodriguez, A. Madabhushi, J. Tomaszeweski, and M. Feldman. Detecting prostatic adenocarcinoma from digitized histology using a multi-scale hierarchical classification approach. In *Engineering in Medicine and Biology Society, 2006. EMBS'06. 28th Annual International Conference of the IEEE*, pages 4759–4762. IEEE, 2006.

[2] C. Gunduz-Demir, M. Kandemir, A.B. Tosun, C. Sokmensuer, et al. Automatic segmentation of colon glands using object-graphs. *Medical image analysis*, 14(1):1, 2010.

[3] K. Jafari-Khouzani and H. Soltanian-Zadeh. Multiwavelet grading of pathological images of prostate. *Biomedical Engineering, IEEE Transactions on*, 50(6):697–704, 2003.

[4] A. Nasser Esgiar, R.N.G. Naguib, B.S. Sharif, M.K. Bennett, and A. Murray. Microscopic image analysis for quantitative measurement and feature identification of normal and cancerous colonic mucosa. *Information Technology in Biomedicine, IEEE Transactions on*, 2(3):197–203, 1998.

[5] M. Roula, J. Diamond, A. Bouridane, P. Miller, and A. Amira. A multispectral computer vision system for automatic grading of prostatic neoplasia. In *Biomedical Imaging, 2002. Proceedings. 2002 IEEE International Symposium on*, pages 193–196. IEEE, 2002.

[6] A. Todman, RNG Naguib, and MK Bennett. Visual characterisation of colon images. In *Proceedings of Medical Image Understanding and Analysis*, pages 161–164, 2001.

[7] A.B. Tosun and C. Gunduz-Demir. Graph run-length matrices for histopathological image segmentation. *Medical Imaging, IEEE Transactions on*, 30(3):721–732, 2011.