Cell-Graph Mining for Breast Tissue Modelling and Classification

C.Çağatay Bilgin [a], Çiğdem Demir [b],Chandandeep Nagi[c], Bülent Yener [a]

[a]Department of Computer Science, Rensselaer Polytechnic Institute, Troy, NY 12180, USA.
[b]Department of Computer Engineering, Bilkent University, Ankara, Turkey.
[c]Mount Sinai Medical Center, NY 10029, USA.

ABSTRACT

Motivation: The most reliable way in the current practice of medicine to diagnose cancer is the pathological examination of a biopsy which has a certain level of subjectivity. To reduce this subjectivity and have a mathematical model for diagnosing cancer tissues we consider the problem of automated cancer diagnosis in the context of breast cancer tissues.

Summary: This work presents a graph theoretical technique that identifies and computes quantitative metrics for tissue characterization and classification. We segmented the digital images of histopatological tissue samples having 10×14 magnification and 960×960 pixels. Then for each image we generated cell-graphs using positional coordinates of cells and surrounding matrix components. These cell-graphs have 500-2000 cells(nodes) with 1000-10000 links depending on tissue and the type of the cell-graph being used. We've calculated a set of global metrics from cell-graphs and used them as the feature set for learning.

Results: We compared our technique with other learning techniques based on intensity values of images, voronoi diagrams of the cells, and the previous technique we proposed for brain tissue images. Among the compared techniques our approach gave %79.1 accuracy whereas we obtained learning ratios of %49.2, %54.1 and %75.9 with intensity based features, voronoi diagrams and our previous technique, respectively.

Contact: bilgic@cs.rpi.edu

1 INTRODUCTION

Breast cancer is the most common cancer and the second leading cause of cancer death among American females, with the current incident rates predicting that 1 in 8 women in the United States will develop breast cancer in her lifetime. Currently, long-term survival is approximately 70%. Diagnosis and staging for prognosis is based on histopathological examination and grading of surgically removed breast tissue and axillary lymph nodes. Prognostic analysis of breast cancer in individual patients currently depends on established clinical, and laboratory parameters such as histopatological grading and hormonal receptor status of individual tumor tissues.

Unfortunately, these parameters are only accurate in approximately 75-80% of the cases, particularly in Stage I tumors. In this group of patients, despite being node negative i.e. tumor confined to the breast with no spread to lymph nodes, 20-30% will recur. Thus, it is important to be able to predict which group of these patients will need chemotherapy to prevent tumor recurrence. Current techniques for diagnosing and predicting the biological behavior of cancer in individual patients are based predominantly on pathological parameters. New molecular techniques are currently being utilized to identify higher risk for specific subgroups of cancer and are in great demand. Unfortunately, reliable prognostic information is still not available in a significant percentage of individuals with common types of cancer, such as breast cancer.

A large set of automated cancer diagnosis tools consists of learning some feature sets. Morphological features such as area, perimeter, and roundness of a nucleus are used in *Esgiar et al*. (1998), *Ganster et al*. (2001), *Glotsos et al*. (2003), *Hamilton et al*. (1987), *Jain et al*. (2004), *Mangasarian et al*. (1995), *Reyes et al*. (1999), *Tasoulis et al*. (2003), *Wolberg et al*. (1995), *Zhou et al*. (2002) for this purpose. Textural features such as the angular second moment, inverse difference moment, dissimilarity, and entropy derived from the co-occurrence matrix are used for diagnosis in in *Esgiar et al*. (1998), *Naguib et al*. (1998), *Glotsos et al*. (2003), *Hamilton et al*. (1997), *Schnorrenberg et al*. (1996), *Tasoulis et al*. (2003), *Wolberg et al*. (1995). To distinguish the healthy and cancerous tissues these systems are trained by using artificial neural networks *Schnorrenberg et al*. (1996), *Tasoulis et al*. (2003), *Zhou et al*. (2002), the k-nearest neighborhood algorithm *Naguib et al*. (1998), *Ganster et al*. (2001), support vector machines *Glotsos et al*. (2003), linear programming *Mangasarian et al*. (1995), logistic regression *Wolberg et al*. (1995), fuzzy *Jain et al*. (2004), and genetic *Reyes et al*. (1999) algorithms. Complimentary to the morphological and textural features, a few of these studies use colorimetric features such as the intensity, saturation, red, green, and blue components of pixels *Ganster et al*. (2001), *Zhou et al*. (2002) and densitometric features such as the number of low optical density pixels in an image *Naguib et al*. (1998), *Hamilton et al*. (1997), *Schnorrenberg et al*. (1996). Another subset of these studies uses fractals that describe the similarity levels of different structures found in a tissue image over a range of scales *Einstein et al*. (1998), *Esgiar, Naguib, Sharif et al*. (2002). These studies use the fractal dimensions as their features and use the k-nearest neighborhood algorithm *Esgiar, Naguib, Sharif et al*. (2002), neural networks, and logistic regression *Einstein et al*. (1998) as their classifiers. Finally, the orientational features are extracted by making use of Gabor filters that respond to contrast edges and line-like features of a specific orientation *Todman et al*. (2001).

There are also some other mathematical diagnosis tools that rely on gene expression (*Ben-Dor et al*. (2000); *Furey et al*. (2000) *Golub et al*. (1999), *Guyon et al*. (2002)) and mass spectroscopy (*Wu et al*. (2003)) to detect a cancer tumor. However, these tools require high technological hard-wired such as micro-arrays (*Guyon et al*. (2002), *Rifkin et al*. (2003)) or mass spectrometers (MALDI, http:info.med.yale.edu/wmkeck/prochem/procmald.htm)

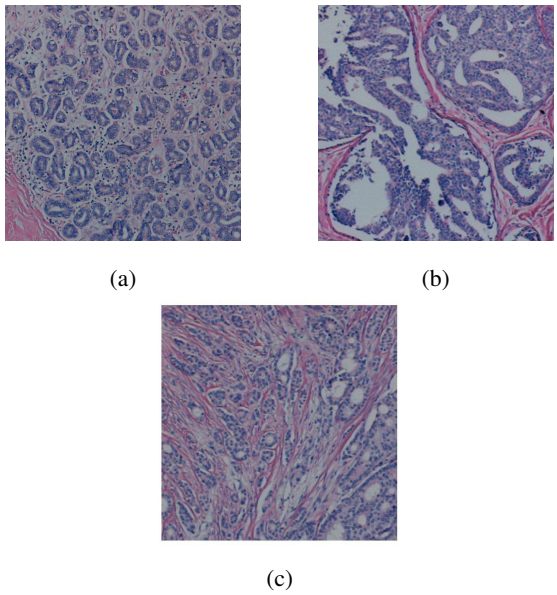
Another approach uses spatial dependency of the cells rather than the intensity values. It constructs a graph of cells from a tissue image and compute graph-theoretical features that quantify how the cells are distributed over the tissue (*Gunduz et al*. (2004); *Weyn et al*. (1999); *Choi et al*. (1997); *Keenan et al*. (2000)). In this approach, a graph of a tissue is defined representing nuclei as vertices and defining the edges as to represent the relationships between adjacent nuclei. In (*Weyn et al*. (1999), *Choi et al*. (1997); *Keenan et al*. (2000)), the Voronoi diagram of the image is constituted and its Delaunay triangulation is built. In these studies the graph-based features are defined on the Delaunay triangulation graph or its corresponding minimum spanning tree. Since the Delaunay triangulation allows the existence of edges between only the adjacent vertices, thus, only the relationships between closely located nuclei are represented in this graph construction method. Moreover, prior to graph construction, this method should carry out the segmentation for each nucleus.

Recently, we generalized the graph based approaches to encode a pairwise spatial relationship between two vertices by constructing cell-graphs (*Gunduz et al*. (2004)). In a cell-graph, the cell clusters of the sample tissue are the vertices and an edge is defined between a pair of these cell clusters based on an assumption that has a biological foundation. For example, if we believe that cells that are spatially close to each other are more likely to interact (e.g., signal) with each other than more distant cells, then a link can be made between them with a probability that decays exponentially with the increasing Euclidean distance between them. Thus links of a cell graph aims at capturing the biological interactions in the underlying tissue.

There are several significant advantages of our cell-graph approach: (1) it enables adapting a rich set of metrics defined by the graph theory to be used as the features, (2) it provides a common framework for both cell level and tissue level feature definition and extraction, and (3) since it uses cell clusters instead of a single segmented cell, it requires only determining the coarse locations of the cells eliminating the necessity of high magnification images. Furthermore, details of a cancerous cell do not need to be resolved, and a specific textural change in the image is not required. Thus the cell-graph approach differs from the previously demonstrated models that also use tissue image analysis, and to the best of our knowledge, it is the first technique to extract the generic organizational principles of cancerous cells from the tissue images for the purpose of both diagnosis and prognostication of cancer tumors.

Applications of graph theory to other problem domains are also impressive. Real-world graphs of varying types and scales have been extensively investigated (*Barabasi et al*. (2002)) in technological (*Shavitt et al*. (2003);*Gunduz and Yener et al*. (2003); *Faloutsos et al*. (1999)), social (*Broder et al*. (2000); *Albert et al*. (1999); *Milgram et al*. (1967); *Newman et al*. (2001); *Wasserman and Faust et al*. (1994); *Liljeros et al*. (2001); *Goldberg et al*. (2003)) and biological systems (*Wuchty et al*. (2003); *Jeong et al*. (2003)). In spite of their different domains, such self-organizing structures unexpectedly exhibit common classes of descriptive spatial (topological) features (*Barabasi et al*. (2002); *Watts and Strogatz et al*. (1998); *Faloutsos et al*. (1999); *Albert et al*. (1999); *Gunduz and Yener et al*. (2003)). These features are quantified by definition of computable metrics. Our computational system is based on the hypothesis that if one can construct a cell-graph then one can define computable metrics on these graphs to identify cancerous tissue and distinguish it from normal and reactive non-neoplastic conditions.

In this paper we present a novel mathematical technique, "mining cell-graphs for modeling and classification of breast tissues". Our modeling approach is based on graph theory which enables us to represet a tissue sample as a graph. We show that cell-graph mining can classify breast tissue samples in different (dis)functional states such as benign, in-situ and invasive.



(a) (b)

(c)

Fig. 1. Microscopic images of tissue samples surgically removed from human breast tissues: **(a)** a benign tissue example, **(b)** an in-situ tissue example, **(c)** an invasive tissue example.

Contributions:

This work is an extension of our previous results on brain tissue images. Because of the underlying architectural differences between the breast and brain tissues, our previous technique is not able to model and classify the samples with more than a %74 accuracy. In this work we extend the cell-graph approach to capture the breast tissue architecture as well.

Organization: The rest of the paper is organized as follows. In section 2 we explain our methodology to generate the cell-graphs of breast tissues. In section 3 we explain the definitions of the metrics that are extracted from our graphs and used as a feature set for learning. We present our experiments and results at section 4, and we conclude our discussion with a summary and a feature perspective for our research.

2 METHODOLOGY

Our technique consists of segmenting the image to extract the cells, modelling the tissue by cell-graphs according to the location of the cells and then learning these graphs using machine learning techniques. Each step is further discussed in the following sections.

2.1 Image Segmentation

1. Segmentation: In order to form cell-graphs on top of the cells, first we need to segment the cells in tissue images. However, image segmentation is still an open question and there are several segmentation techniques that are proposed for different types of images. K-means algorithm, which clusters the pixels of images according to their RGB values into clustering vectors, gave satisfactory results for breast tissue images. The clustering vectors obtained from K-means algorithm are assigned as either being cell or background by a human expertise. The clustering vectors are estimated as to minimize the following error function E,

$$E = \sum_{j=1}^{k} \sum_{x_n \in Sj} (x_n - \mu_j)^2$$

   This step is depicted as the transition from figure 2a to 2b.

2. Node Identification: The next step is to translate the class information to node information. After the image segmentation, we have the pixels that constitute a cell but still boundaries of the cells are not available. We've placed a grid on the resulting images of image segmentation to identify the cells. For each grid entry we've calculated the probability of being a cell as the ratio of cell pixels to the total number of pixels in the grid. Then we've applied thresholding to decide whether this grid entry is a cell or not.

   Note that there are two parameters in this step, namely grid size and threshold value. The grid size depends on the actual cell size and therefore should be decided independent from the rest of the work. Increasing the threshold value will help to eliminate noise in the image segmentation but increasing it beyond an optimum value will result in the loss of cell information. Therefore, we need a threshold value which can identify the cells and eliminate the noise in the image. This step can also be considered as downsampling of the image. The result of node identification is given in figure 2d.

2.2 Cell-Graph Generation

After the image segmentation, we have the locations of the cells, which are the centers of the grid entries. We build our graphs on top of these grid entries. Formally a graph is represented by $G = (V, E)$ where V is the vertex set of the graph and E is the edge set of the graph. After image segmentation step we have the vertex set of the graphs and in cell-graph generation we form the edges of the graphs.

We've constructed three different kinds of cell-graphs capturing the pairwise distance relationship between the nodes. These three different kinds of cell-graphs are explained in the following sections.

2.2.1 Simple Cell-Graphs: In simple cell-graphs we set a link between two nodes if the euclidean distance is less than a threshold. The euclidean distance between two cells is given by

$$d(u, v) = \sqrt{(u_x - v_x)^2 + (u_y - v_y)}.$$

where u_x and u_y are x and y coordinates of node u respectively.

These graphs form a relation between nodes if they are close to each other.

2.2.2 Probabilistic Cell-Graphs: The probabilistic model is a more general version of simple cell-graphs. In this model we build a link between two nodes with a certain probability which is given by given by

$$P(u, v) = d(u, v)^{-\alpha} \text{ for nodes u and v.}$$

That is, these graphs may build links between two nodes which are far away from each other. Note that probabilistic graphs do not necessarily form links between two nodes even if the distance between two nodes is small. But still, it's more likely that nodes that are close to each other will have links and nodes that are far away from each other will not have links between them.

2.2.3 Hierarchical Cell-Graphs: The previous two forms of graphs capture the global distribution of the cells and were usefull for brain tissue images. However, the underlying architectures of brain and breast tissues are different from each other. Breast tissues have lobular architecture whereas, brain tissues do not have such higher level structures. For breast tissues, the pairwise relations of cells within the same gland as well as the pairwise relations of different glands are therefore important. To capture the lobular architecture of the breast tissues we need an hierarchical representation of the tissues. We formed our hierarchical graphs similar to the way we formed our cell-graphs. After the node identification step we had our nodes (cells) of the graphs. In order to find the clusters (lobs) of the tissues, we placed a grid on top of these cells. We calculated the number of cells in the grid and calculated the probability of being a cluster for each grid entry. Then we formed our graphs on these clusters. We can think this step as further downsampling of the image to capture the cell clusters. This step is shown in figure 2e.

Note that the presence of a link between nodes does not specify what kind of relationship exists between the nodes (cells); it simply indicates that a relationship of some sort is proposed to exist, and that it is dependent on the distance between cells. Surprisingly, this measure alone is sufficient to reveal important, diagnostic structural differences in human tissues.

2.3 Cell-Graph Mining

In order to learn the differences between the graphs we need to find a way to extract the properties (metrics) of these graphs. The metrics that are computed for each graph are explained in section 3. After calculating our metrics, we have scaled our data since some metrics are too large and some of them are too small therefore effecting the learning significantly. We've scaled each metric to the range $[-1, 1]$ for a better comparasion.

We have used support vector machines (SVM) as our main classifier. The SVM algorithm creates the optimal separating hyperplane between data points such that the data points of different classes fall into the opposite sides of this hyperplane. If there is no hyperplane that separates these two classes (i.e., if the data is not linearly separable), this algorithm creates a hyperplane that leads to the least error.

Parameters of the optimal separating hyperplane are derived by solving a quadratic programming optimization problem with linear equality and inequality constraints; this optimization problem maximizes the margin. In case of a nonseparable data set, the slack variables are introduced to minimize the error. An important feature of support vector machines is the use of kernel functions. The kernel function transforms the input space to a new space and

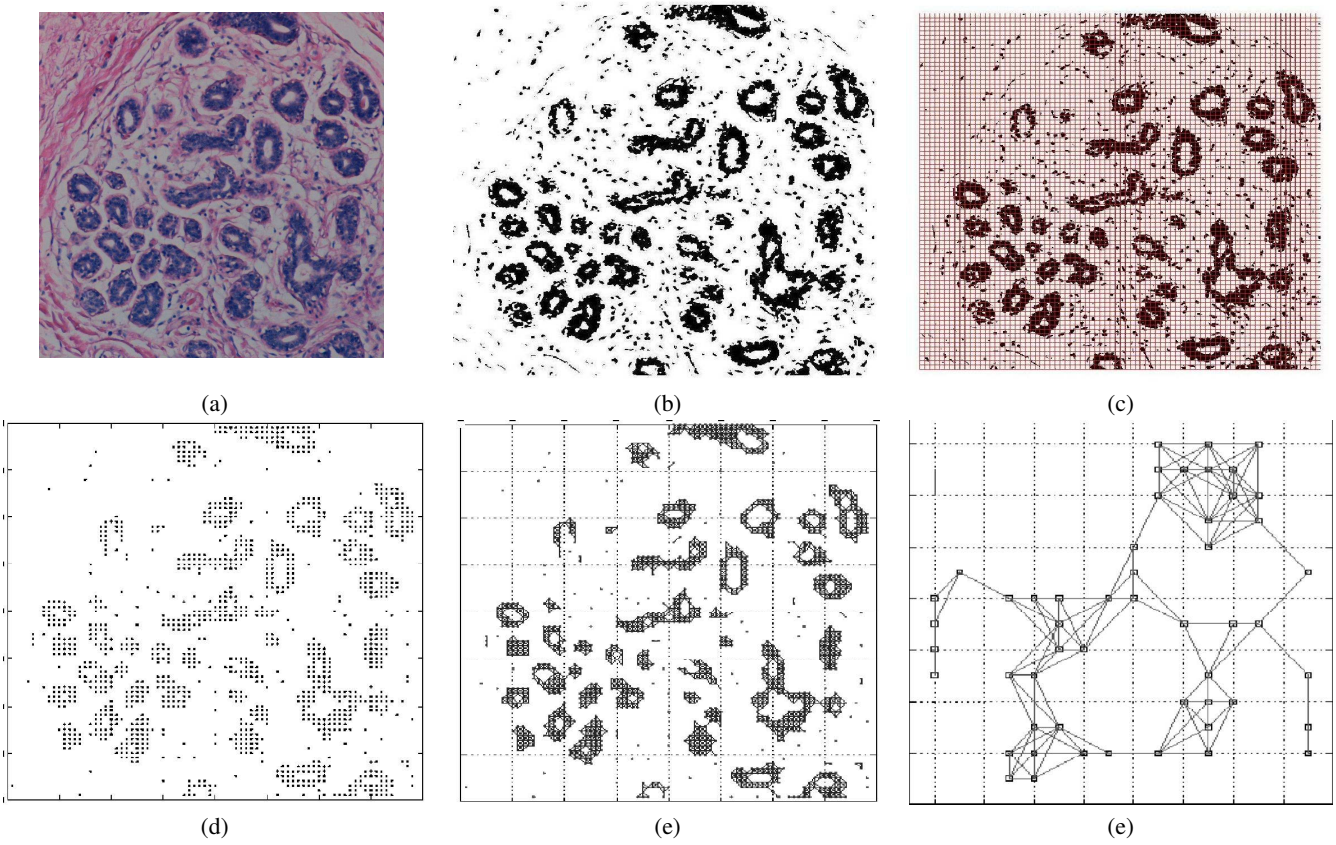


Fig. 2. The steps of our methodology. (a) Original tissue image is opened in RGB space. (b) The result of k-means segmentation, black points are part of cells and white points are treated as background. (c) The application of grid and thresholding to the resulting segmentation. Appling a thresholding will get rid of the noise in the segmentation and the center of grid entries will be used as the locations of cells. (d) The overall result of node identification. (e) Simple cell-graphs are formed based on the location information of the cells. (f) A bigger grid is applied to the image to capture the cell clusters. Each grid entry is then thresholded to get the clusters. After cluster identification, hierarchical graphs are build on cluster cells.

allows the algorithm to find the optimal separating hyperplane in this new space. The use of nonlinear kernel functions allows using non-linearity without explicitly requiring a non-linear algorithm.

We've used SVMs with the radial basis kernel

$$K(x, y) = e^{\gamma \|x-y\|^2}.$$

To find the best parameters of C where C is the penalty parameter and γ, we've used cross-validation. After finding C and γ we trained our training set with these parameters. We then tested our classifier on the test set which is completely different than the training set.

3 METRICS

In order to have a quantitive representation of the graphs we extracted some metrics from the graphs. We use several different topological properties defined on the entire graph (i.e. global graph metrics). These cell-graph features are as simple as the number of neighboring cells which corresponds to the degree of a node.

1. The simplest metric is the **number of nodes** in the graph. The degree of a node is defined as the number of its edges. Using the distribution of the node degrees, we compute the **average degree** as a global metric. A cancerous cell cluster or tissue has typically larger value for this metrics. On the other hand, it is not always the indicator for cancer as in the case of in-situ cell clusters or tissues.
2. Another graph metric is the **clustering coefficient** of a node C_i, which is defined as $C_i = (2E_i)/(k(k+1))$, where k is the number of neighbors of the node i and E_i is the number of existing links between its neighbors (Dorogovtsev and Mendes, 2002). This metric quantifies the connectivity information in the neighborhood of a node. We use the average clustering coefficient as a global metric.
3. The **path length** between two nodes is defined as their shortest path length in the graph, taking the weight of each link as a unit length.
4. Given shortest path lengths between a node i and all of the reachable nodes around it, the **eccentricity** and the **closeness** of the node i are defined as the maximum and the average of these shortest path lengths respectively. The maximum value of the eccentricity, also known as the **diameter** of a graph, is another metric for the classifier. This set of metrics reflects the centrality of the node. We conjecture that their smaller values

indicate that the node is close to the center of the cell-graph, and hence, to the center of the cancer invasion.

5. **Central points** of the graph is defined as the points having an eccentricity equal to the radius. We've used this metric for the learning as well.
6. The hop plot value reflects the size of a neighborhood between any two nodes within a hop. For hop h, the hop plot value is defined as the number of node pairs such that the path length between these node pairs is less than or equal to h hops. Using the hop plot value distributions, two global features are computed. The first one is the **hop-plot exponent**, which is computed as the slope of the hop plot values as a function of h in log-log scale. The second global feature is the **effective diameter**, which is defined as $\varepsilon = \frac{N^2}{(N+2E)^{\frac{1}{H}}}$ where N and E are the number of nodes and edges, and H is the hop plot exponent.
7. We've also computed some global graph metrics which are not directly computed from the distributions of the local graph metrics. For example the ratio of the size of the giant connected component over the size of the entire graph is one of the distinguishing features in the learning step. In graph theory, the **giant connected component** of a graph is defined as the largest set of the nodes where all of the nodes in this set are reachable from each other.
8. Other global graph metrics are the **percentages of the isolated and the end nodes** in the entire graph. A node of a graph is called isolated point if it has no edges, i.e., if it has a degree of 0. A node of a graph is called end point if it has only one edge.

4 EXPERIMENTS

4.1 Data Set Preparation

The tissues are randomly selected from the archived Mount Sinai School of Medicine (MSSM) Pathology Department archives. For each subject, a group of representative slides are chosen by the pathologist, a subject identifier (i.e. the accession number on slide labels) are removed after diagnostic tabulation in a coded manner. Then, the coded data are kept and, hence, there is no any direct linked back to the subjects. These cases are reviewed by breast pathologist Dr. Nagi in collaboration with Shabnam Jaffer MD. at MSSN to reach a consensus.

This selection is made uniformly and randomly, but preference are given to cases from the last 5 years, unless an adequate number of available cases are not reached. This allows access to more cases that have been worked up and managed with modern clinical, radiological, surgical and pathology techniques. All patient populations, regardless of age, sex, or race, are be included. Patient reports are available to the pathologist on a pathology database. First selection is performed based on diagnostic categories, such as *all patients diagnosed with invasive duct carcinoma from 1999 to date*. After initial selection, individual cases are examined under the microscope to confirm the diagnosis, and technical adequacy of the material. This is performed by two independent pathologists to further ensure reliability and accuracy. After a glass slide is chosen for the study, it will be numerically coded, and patient identifiers will be removed. The coded tally of individual cases is secured in the pathologist's office. Digital photomicrographs are obtained in a standardized fashion with regards to magnification and illumination.

Three major diagnostic groups are be formed and analyzed. The first group consists of normal breast tissues. These are obtained from surgical pathology material. The second group consists of benign reactive processes, such as hyperplasia, radial scar or inflammatory changes. Florid hyperplasia may simulate duct carcinoma in situ based on cellularity. Histopathologically, however, they are usually easily discerned from neoplasms. The rationale for including this category is to test the computer algorithms, and prove that high cellularity alone is not mistaken for a neoplastic process using the model that is proposed. Other benign conditions such as sclerosing adenosis will also be tested on the computer model to ensure that a low power pattern is not confused with invasive carcinoma. The third group is infiltrating carcinomas. The definition and grading of these tumors is performed according to the published guidelines of the modified Bloom Richardson criteria.

We conduct our experiments on the data set that comprises the images of cancerous and benign breast tissues. This data set consists of both invasive and noninvasive (ductal carcinoma in situ [DCIS]) cancerous tissues. Similar to the brain tissue data set, this data set contains the tissues of patients that were randomly chosen from the Pathology Department archives in Mount Sinai School of Medicine and each of these tissues was stained with hematoxylin and eosin technique. A Nikon Coolscope Digital Camera/Scanner was used to take the images of breast tissue samples. Images were taken in the RGB color space prior to color quantization. The magnification of images is 100×14 and they are taken by using a 10 microscope objective lens and there is another lens at the eye end.. In our experiments, we use tissue images with a resolution of 960×960.

Our data set contains images of 446 breast tissue samples that are removed from 36 different patients. We split this data set into the training and test sets each of which consists of 18 patients; the patients of the training and test sets are completely different. In this data set, some patients have tissue samples of more than one tissue type (for example, the same patient might have both invasive cancerous and benign tissue samples). In the training set, we use 84 invasive cancerous tissue images of 10 patients, 38 non-invasive cancerous (DCIS) tissue images of 5 patients, and 82 benign tissue images of 10 patients.

In the test set, the tissue and patient distribution is as follows: 118 invasive cancerous tissue images of 9 patients, 55 DCIS tissue images of 6 patients, and 69 benign tissue images of 9 patients.

4.2 Comparative Results

We have calculated the accuracy of our learning technique and then compared it against the intensity based approach, voronoi based approach, simple cell graphs, and hierarchical cell-graphs.

In the intensity-based approach features are extracted from the gray-level or color histogram of pixels and do not provide any information about the spatial distribution of pixels. At the cellular level, the intensity histogram of pixels surrounded by the boundary of a nucleus is employed to define features. For example, using gray-level histograms, Weyn et al. computed the sum and mean of the optical densities of the pixels located in a nucleus and defined these values as the intensity-based features of the nucleusWe extracted intensity-based features by employing the RGB values of pixels in a tissue. For each color channel, we computed the mean and standard deviation of pixel values in an entire image and used these values as the feature set of the classifier.

Table 1. Hierarchical Cell-Graph Results

Link Threshold	Grid Size 4	5	8	10	16
1	60.1	64.3	68.9	76.4	69.5
2	67.0	66.0	65.0	81.8	68.0
3	59.6	70.0	73.9	75.9	70.0
4	57.6	60.6	74.9	70.0	69.5
5	68.5	66.5	70.4	69.5	69.5
6	61.6	60.6	65.0	70.4	69.5
7	64.0	58.6	64.0	66.0	69.5
8	60.6	71.4	63.1	66.0	69.5
9	58.6	57.6	63.5	66.0	69.5
10	54.2	53.7	65.5	66.0	69.5

The choice of grid size for clusters affects the learning ratio significantly. We obtain the best results when the grid size is 10 and link threshold is 2. On the other hand, we have a learning ratio of %54.2 when the grid size is 4 and the link threshold is 10.

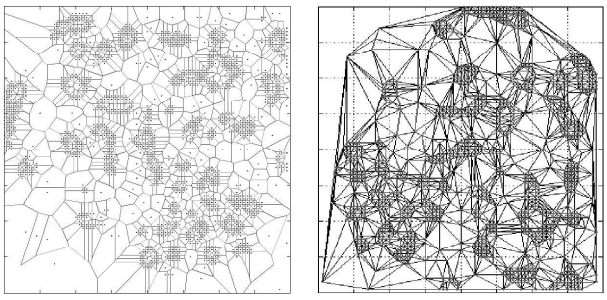
Table 2. Probabilistic Cell-Graphs

Link Threshold	5	6	7	8	9
Benign	92.0±0.03	88.7±0.04	89.2±0.04	91.6±0.02	91.1±0.03
InSitu	50.9±0.04	54.9±0.06	55.1±0.05	50.2±0.04	47.8±0.07
Invasive	79.2±0.04	75.9±0.04	74.6±0.07	77.1±0.04	78.1±0.03
Overall	74.5±0.02	73.2±0.03	72.6±0.04	73.1±0.01	72.9±0.02

Table 3. Comparison of the techniques

	Intensity	Delaunay	Probabilistic	Simple	Hierarchical	Hybrid
Benign	20.6	80.9	90.5	84.7	84.9	90.9
InSitu	69.1	16.3	51.8	51.6	69.9	57.3
Invasive	54.6	56.3	77.0	85.6	80.7	86.3
Overall	49.1	54.1	73.4	75.8	74.0	79.2

In order to quantify the spatial distribution of nuclei, Voronoi diagrams and their Delaunay triangulations are proposed On a tissue image, the Voronoi diagram constitutes convex polygons for each nucleus. For a particular nucleus, every point in its polygon is closer to itself than to another nucleus in the tissue. The dual graph of the Voronoi diagram is the Delaunay triangulation. The Voronoi diagram of a sample tissue image and its Delaunay triangulation are illustrated in figure 3. In this approach, we define the Voronoi diagram on cell-clusters that we identify in node identification step. Then we evaluate the metrics explained in section 3 for these diagrams. These metrics are then given as the feature set to the classifier.



(a) (b)

Fig. 3. (a)The voronoi cells of the tissue. (b)The dual of the voronoi diagram, Delaunay triangulation.

In table 1 we see that increasing the link threshold value also increases the learning ratio up to some point. Increasing the link threshold beyond this value decreases the learning ratio.

In table 3 we give the comparative results of the techniques discussed in the paper. The intensity based approach achieves a learning ratio of 49.2 which is the worst ratio amongst the the others. Delaunay triangulation of the cells gave better results than the intensity based approach since this technique embeds the spatial distribution of the cells in learning. Simple cell-graphs, however, embeds the spatial distribution of the cells better than the Delaunay triangulation and achieves a %75.93±2.53 learning ratio on everage for link threholds varying between 1 and 10. Probabilistic cell-graphs doesn't change the result significantly compared with the simple cell-graphs and they achieve a learning ratio of %73.4±1.24. The learning ratio of hierarchical graphs is dependent on the choice of the grid size and the link threshold. A good choice of these metrics is small link thresholds and a fairly big to find the clusters. For hierarchical graphs after some point increasing the link threshold doesn't change the learning ratio, since we obtain a complete graph where each node has a link to the other nodes. Therefore when we are calculating the average value for the hierarchical cell-graphs we omitted the link thresholds greater than 7. We've used a grid size of 10 which is able to capture the cell clusters. The learning ratio for hierarchical graphs is %74±4.89. As the last technique we've combined the intensity features, the metrics calculated from simple cell-graphs and hierarchical cell-graphs and used this set as the feature set of our classifier. This hybrid approach is calculated for a grid size of 10 and the average value for this technique is %79.1.

5 CONCLUSION

Previously, we used cell-graphs to model and classify brain tissue samples which present a diffusive structure. In this work we extend and enhance the cell-graph approach to modeling and classification of breast tissue samples which has a lobular/glandular architecture, thus differ from brain tissues significantly.

Cell-graphs enable us to identify and compute a rich set of features that represent the two dimensional structure information of breast tissues. The feature sets are input to a support vector machine for classification of benign, invasive and noninvasive (ductal carcinoma in situ) cancerous tissues. We show that accuracy of classification depends significantly to the construction of cell-graphs

which needs to capture the characteristics of underlying native tissue. A computational comparison of our approach to the related work in the literature shows that cell-graphs are much more accurate. However, we believe that accuracy can be improved further by increasing the data size and image segmentation.

Finally we note that cell-graphs can be used as a decision support system by pathologists, or simply for training.

ACKNOWLEDGMENT

The authors thank to Shabnam Jaffer MD at Mount Sinai School of Medicine for her helps.

REFERENCES

A. Ben-Dor, L. Bruhn, N. Friedman, I. Nachman, M. Schummer, and Z. Yakhini (2000) Tissue classification with gene expression profiles, *Comput Biol*. 7(3-4):559-583.

I. Guyon, J. Weston, S. Barnhill, and V. Vapnik (2002) Gene selection for cancer classification using support vector machines, *Machine Learning*, 46(1-3):389-422.

A.N. Esgiar, R.N. Naguib, M.K. Bennett, and A. Murray (1998) Automated Feature Extraction and Identification of Colon Carcinoma, *J. Analytical and Quantitative Cytology and Histology*, vol. 20, no. 4, pp. 297-301.

H. Ganster, P. Pinz, R. Rohrer, E. Wildling, M. Binder, and H.Kittler. Automated Melanoma Recognition (2001), *IEEE Trans. Medical Imaging*, vol. 20, no. 3, pp. 233-239.

D. Glotsos, P. Spyridonos, P. Petalas, G. Nikiforidis, D. Cavouras, P. Ravazoula, P. Dadioti, and I. Lekka (2003) Support Vector Machines for Classification of Histopathological Images of Brain Tumour Astrocytomas, *Proc. Intl Conf. Computational Methods in Sciences and Eng.*, pp. 192-195.

P.W. Hamilton, D.C. Allen, P.C. Watt, C.C Patterson, and J.D. Biggart (1987) Classification of Normal Colorectal Mucosa and Adenocarcinoma by Morphometry, *Histopathology*, vol. 11, no. 9, pp. 901-911.

R. Jain and A. Abraham (2004) A Comparative Study of Fuzzy Classification Methods on Breast Cancer Data, *Australiasian Physical and Eng. Sciences in Medicine*.

O.L. Mangasarian, W.N. Street, and W.H. Wolberg (1995) Cancer Diagnosis and Prognosis via Linear Programming, *J. Operational Research*, vol. 43, no. 4, pp. 570-577.

C.A. Pena-Reyes and M. Sipper (1999) A Fuzzy Genetic Approach to Breast Cancer Diagnosis, *Artificial Intelligence in Medicine*, vol. 17, no. 2, pp. 131-155.

D.K. Tasoulis, P. Spyridonos, N.G. Pavlidis, D. Cavouras, P. Ravazoula, G. Nikiforidis, and M.N. Vrahatis (2003) Urinary Bladder Tumor Grade Diagnosis Using On-Line Trained Neural Networks, *Proc. Knowledge Based Intelligent Information Eng. Systems Conf*, pp. 199-206.

W.H. Wolberg, W.N. Street, D.M. Heisey, and O.L. Mangasarian (1995) Computer-Derived Nuclear Features Distinguish Malignant from Benign Breast Cytology, *Human Pathology*, vol. 26, no. 7, pp. 792-796.

Z.H. Zhou, Y. Jiang, Y.B. Yang, and S.F. Chen (2002) Lung Cancer Cell Identification Based on Artificial Neural Network Ensembles, *Artificial Intelligence in Medicine*, vol. 24, no. 1, pp. 25-36, 2002.

A.N. Esgiar, R.N.G. Naguib, B.S. Sharif, M.K. Bennett, A. Murray (1998) Microscopic Image Analysis for Quantitative Measurement and Feature Identification of Normal and Cancerous Colonic Mucosa, *IEEE Trans. Information Technology in Biomedicine*, vol. 2, no. 3, pp. 197-203.

P.W. Hamilton, P.H. Bartels, D. Thompson, N.H. Anderson, and R. Montironi (1997) Automated Location of Dysplastic Fields in Colorectal Histology Using Image Texture Analysis, *J. Pathology*, vol. 182, no. 1, pp. 68-75.

F. Schnorrenberg, C.S. Pattichis, C.N. Schizas, K. Kyriacou, and M.Vassiliou (1996) Computer-Aided Classification of Breast Cancer Nuclei, *Technology and Health Care*, vol. 4, no. 2, pp. 147-161.

A.J. Einstein, H.S. Wu, M. Sanchez, and J. Gil (1998) Fractal Characterization of Chromatin Appearance for Diagnosis in Breast Cytology, *J. Pathology*, vol. 185, pp. 366-381.

A.N. Esgiar, R.N.G. Naguib, B.S. Sharif, M.K. Bennett, A. Murray (2002) Fractal Analysis in the Detection of Colonic Cancer Images, *IEEE Trans. Information Technology in Biomedicine*, vol. 6, no. 1, pp. 54-58.

A.G. Todman, R.N.G. Naguib, and M.K. Bennett (2001) Orientational Coherence Metrics: Classification of Colonic Cancer Images Based on Human Form Perception, *Proc. Canadian Conf. Electrical and Computer Eng*, vol. 2, pp. 1379-1384,

Furey,T.S., Christianini,N., Duffy,N., Bednarski,D.W.,Schummer,M. and Hauessler,D. (2000) Support vector machine classification and validation of cancer tissue samples using microarray expression data, *Bioinformatics*, 16, 906914.

Golub,T.R., Slonim,D.K., Tamayo,P., Huard,C., Gaasenbeek,M.,Mesirov,J.P., Coller,H., Loh,M.L., Downing,J.R., Caligiuri,M.A., Bloomfield,C.D. and Lander,E.S. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, *Science*, 286, 531537.

Guyon,I., Weston,J., Barnhill,S. and Vapnik,V. (2002) Gene selection for cancer classification using support vector machines, *Machine Learning*, 46, 389422.

Wu,B., Abbott,T., Fishman,D., McMurray,W., Mor,G., Stone,K., Ward,D., Williams,K. and Zhao,H. (2003) Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data. *Bioinformatics*, 19, 16361643.

Rifkin,R., Mukherjee,S., Tamayo,P., Ramaswamy,S., Yeang,C.-H., Angelo,M., Reich,M., Poggio,T., Lander,E.S., Golub,T.R. and Mesirov,J.P. (2003) Ananalytical method for multiclass molecular cancer classification, *SIAM Rev*, 45, 706723.

Gunduz, C., B. Yener, and S. H. Gultekin (2004) The cell graphs of cancer, *Bioinformatics*, 20: i145-i151.

Weyn B., G. Van de Wouwer, S. Kumar-Singh, A. Van Daele, P. Scheunders, E. Van Marck, and W. Jacob (1999) Computer-assisted differential diagnosis of malignant mesothelioma based on syntactic structure analysis, *Cytometry*, 35:23-29.

Choi H.-K., T. Jarkrans, E. Bengtsson, J. Vasko, K. Wester, P.-U. Malmstrom, and C. Busch (1997) Image analysis based grading of bladder carcinoma. Comparison of object, texture and graph based methods and their reproducibility, *Anal Cell Pathol*, 15:1-18.

Keenan S. J., J. Diamond, W. G. McCluggage, H. Bharucha, D. Thompson, B. H. Bartels, and P. W. Hamilton (2000) An automated machine vision system for the histological grading of cervical intraepithelial neoplasia (CIN), *J Pathol*, 192(3):351-362.

Barabasi A-L (2002) Linked: The New Science of Networks, *Perseus Books Group; 1ST edition* .

Shavitt Y., X. Sun, A. Wool, and B. Yener (2003) Computing the unmeasured: an algebraic approach to Internet mapping, *IEEE Journal on Selected Areas in Communications*, 22(1):67-78.

Gunduz C. and B. Yener (2003) Accuracy and sampling trade-offs for inferring Internet router graph, *Rensselaer Polytechnic Institute, Department of Computer Science, TR-03-09*.

Faloutsos M., P. Faloutsos, and C. Faloutsos (1999) On power-law relationships of the Internet topology, *in Proceedings of ACM/SIGCOMM*, 251-262.

Broder A., R. Kumar, F. Maghoul, P. Raghavan, and R. Stata (2000) Graph structure in the Web, *Proceedings of the 9th International World Wide Web Conference*, 247-256.

Albert R., H. Jeong, A.-L. Barabasi (1999) Diameter of the World-Wide Web, *Nature*, 401:130-131.

Milgram S (1967) The small-world problem, *Psychology Today*, 1967, 2:61-67.

Newman M. E. J. (2001) Who is the best connected scientist? A study of scientific coauthorship networks, *Physics Review*, E64.

Wasserman S. and K. Faust (1994) Social network analysis: methods and applications, *Cambridge University Press*.

Liljeros F., C. R. Edling, L. A. N. Amaral, H. E. Stanley, and Y. Aberg (2001) The web of human sexual contacts, *Nature*, 411:907-908.

Goldberg M., P. Horn, M. Magdon-Ismail, J. Riposo, D. Siebecker, W. Wallace, B. Yener (2003), Statistical modeling of social groups on communication networks *First conference of the North American Association for Computational Social and Organizational Science (CASOS 03)*.

Wuchty S., E. Ravasz, and A.-L. Barabasi (2003) The architecture of biological networks, *in T.S. Deisboeck, J. Yasha Kresh and T.B. Kepler (eds.), Complex Systems in Biomedicine, Kluwer Academic Publishing*.

Jeong, B. Tombor, R. Albert, Z. N. Oltvai, A.-L. Barabasi (2000) The large-scale organization of metabolic networks, *Nature* 407:651-654.

Watts D. and S. Strogatz (1998) Collective dynamics of small-world networks, *Nature* 393:440-442.