Cell-Graph Mining for Breast Tissue Modeling and Classification

Cagatay Bilgin[a], Cigdem Demir[b],Chandandeep Nagi[c], Bulent 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— **The most reliable way to diagnose cancer in the current practice of medicine is through 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 tissues. In this work we present graph theoretical techniques that identify and compute quantitative metrics for tissue characterization and classification. We segment the digital images of histopatological tissue samples using k-means algorithm. For each segmented image we generate different 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 the tissue and the type of cell-graph being used. Having generated the graphs, we calculate a set of global metrics from cell-graphs and use them as the feature set for learning. We compare our technique, hierarchical cell graphs, with other techniques based on intensity values of images, Delaunay triangulation of the cells, the previous technique we proposed for brain tissue images and with the hybrid approach that we introduce in this paper. Among the compared techniques, hierarchical-graph approach gives 81.8% accuracy whereas we obtain 61.0%, 54.1% and 75.9% accuracy with intensity-based features, Delaunay triangulation and our previous technique, respectively.**

I. Introduction

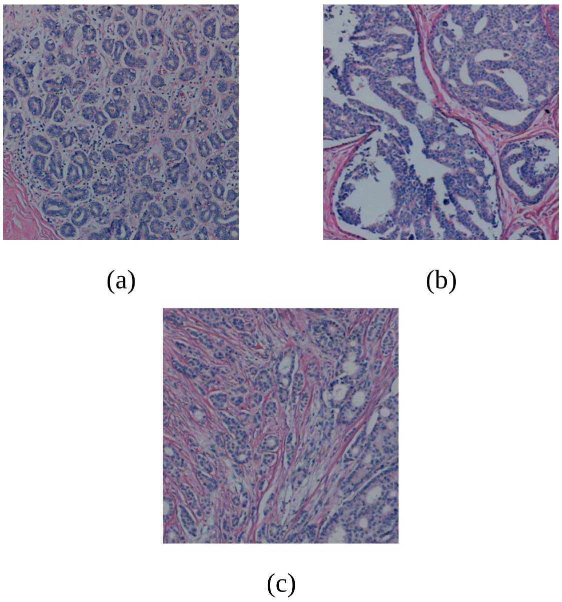
Breast cancer is the most common cancer and the second leading cause of cancer death among American females. The current incident rates predict that 1 in 8 women in the United States will develop breast cancer in their lifetime. Currently, long-term survival is approximately 70%. Early diagnosis of breast cancer is crucial and the 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 exists in literature which are based on learning some feature sets. Morphological features such as area, perimeter, and roundness of a nucleus are used in [7], [22], [12], [29], [31], [28], [10], [13], [38], [41] 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 [7], [6], [12], [30], [14], [13], [38]. To distinguish the healthy and cancerous tissues these systems are trained by using artificial neural networks [14], [13], [41], the k-nearest neighborhood algorithm [6], [22], support vector machines [12], linear programming [28], logistic regression [38], fuzzy [31], and genetic [10] 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 [22], [41] and densitometric features such as the number of low optical density pixels in an image [6], [30], [14]. 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 [3], [5]. These studies use the fractal dimensions as their features and use the k-nearest neighborhood algorithm [5], neural networks, and logistic regression [3] 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 [2].

There are also some other mathematical diagnosis tools that rely on gene expression [1], [16], [18], [21] and mass spectroscopy [39] to detect a cancer tumor. However, these tools require high technological hard-wired such as microarrays [21], [32] or mass spectrometers [42].

There are other approaches using spatial dependency of the cells rather than the intensity values. These approaches construct a graph of cells from a tissue image and compute graph theoretical features that quantify how the cells are distributed over the tissue [20], [36], [11], [24]. In these approaches, a graph of a tissue is defined by representing nuclei as vertices and defining edges to capture relationships between nuclei. In [36], [11] and [24], 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, only the relationships between closely located nuclei are represented in this method. Moreover, prior to graph construction, this method should carry out the segmentation for each nucleus.



(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: In this paper we try to capture pathologists' rule of thumbs. We model the tissues based on graph theory and show that cell-graph mining can classify breast tissue samples in different (dis)functional states such as benign, in-situ and invasive. This paper extends our previous work on brain tissues. There is an underlying architectural difference between breast and brain tissue examples, therefore, a new set of features and new graph based modeling techniques are needed. In this work we present different graph based approaches, such as hierarchical graphs and hybrid based learning and different sets of features to differentiate between cancerous, benign and in-situ tissues.

Organization: The rest of the paper is organized as follows. In section 2 we explain our methodology for generating cell-graphs of breast tissues. In section 3 we explain the definitions of the metrics which are extracted from our graphs and used as feature sets for learning. We present our experiments and results in section 4, and we conclude our discussion in section 5.

II. Methodology

Our technique consists of segmenting the image to extract the cells, modelling the tissue by 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.

A. *Image Segmentation*

1) Segmentation: In order to form 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 are estimated as to minimize the following error function E,

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

where μ_j is the center of jth cluster and x_n's are the intensity values of the images. 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. The image segmentation procudes pixels that constitute a cell but still the boundaries of the cells are not available. We placed a grid on the resulting images of segmentation to identify the cells. For each grid entry we calculated the probability of being a cell as the ratio of cell pixels to the total number of pixels in the grid. Then we 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 considered 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.

B. *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 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.

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

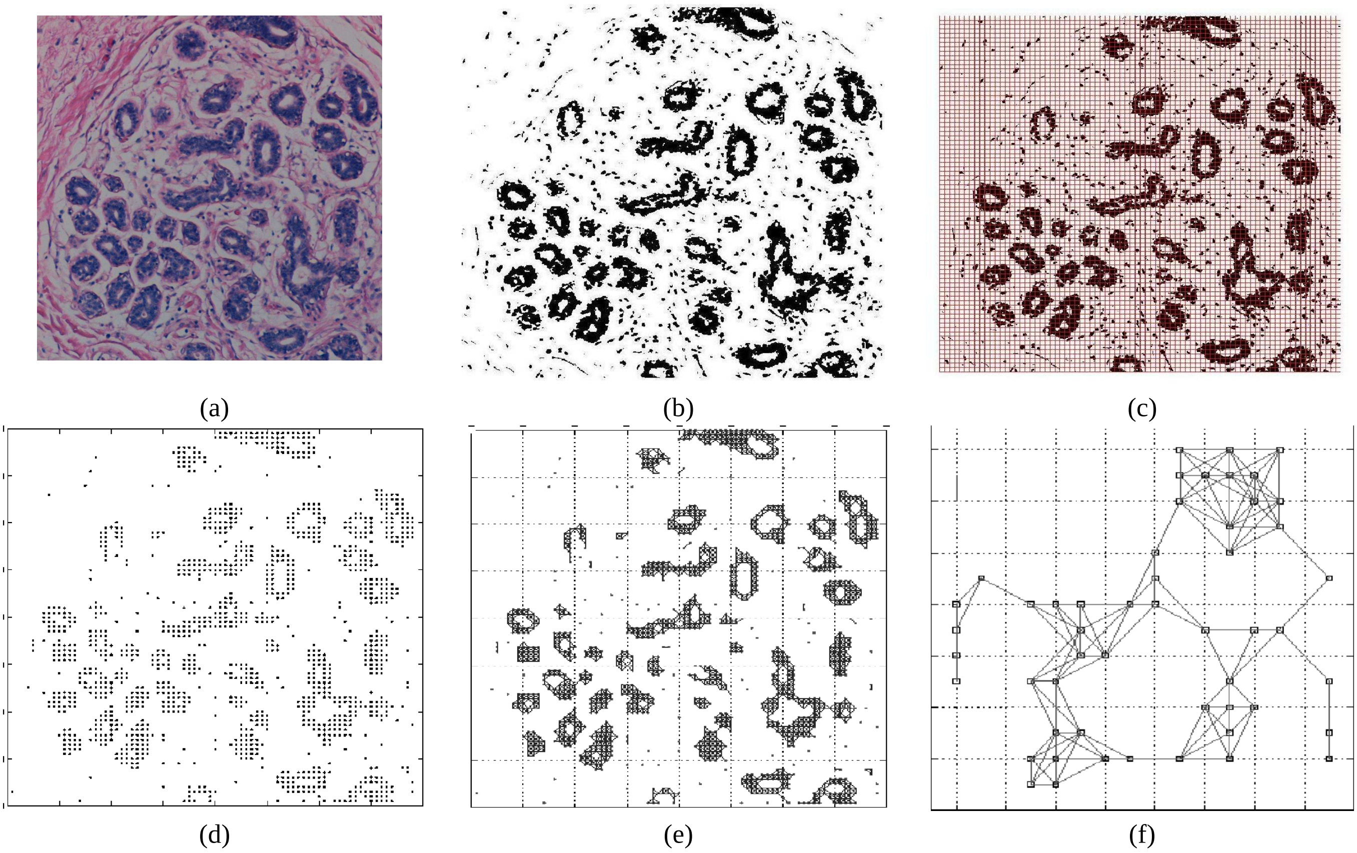


Fig. 2. The steps of our methodology. (a) Original tissue image is opened in RGB space. (b) The result of k-means segmentation, black points are part of cells and white points are treated as background. (c) The application of grid and thresholding to the resulting segmentation. Appling a thresholding 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.

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

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) 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

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

This kind of graphs introduce building a link between two nodes which are not neccessarily close neighbors. d is defined as the euclidean distance between the nodes as in the case of simple cell-graphs. The use of α introduces a decaying property for building links. As the distance between nodes increases, the probability of linking them decreases. Note that probabilistic graphs do not necessarily form links between two nodes even if the distance between the nodes is small. Yet, it is more likely for the nodes that are close to each other will be linked while the nodes that are farther away will not be linked.

3) Hierarchical Cell-Graphs:: The previous two forms of graphs capture the global distribution of the cells and were particularly useful for brain tissue images. However, there is an underlying architectural difference between the brain and breast tissues. Breast tissues have lobular architecture whereas brain tissues do not have such higher level structures. For breast tissues, the pairwise relationship of cells within the same gland as well as 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 (lobes) of the tissues, we placed a grid on top of these cells. We calculated the number of cells in the grid and by dividing this number by the grid size, we calculated the probability of being a cluster for each grid entry. After obtaining the probability values for each grid entry, we set a threshold value and considered the grid entries with a probability greater than this threshold as a cluster. We then formed our graphs on these clusters. This step can actually be considered as further downsampling of the image to capture the cell clusters as depicted 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.

C. 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 III. After calculating our metrics prior to learning, the metrics are scaled since some metrics are too large and some of them are too small therefore effecting the learning significantly. We 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 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 have used SVM classifier with a radial basis kernel

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

We applied grid search to find out the best parameters for the SVM. After finding these parameters we trained our training set with these parameters and then tested it on the test set which was disjoint to the training set.

III. Metrics

In order to have a quantitative 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 values for these 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.
5) **Central points** of the graph is defined as the points having an eccentricity equal to the radius. We 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 $\varepsilon = \dfrac{N^2}{(N+2E)^{\frac{1}{H}}}$ where N and E are the number of nodes and edges, and H is the hop plot exponent.
7) We 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.

IV. Experiments

A. 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 first chosen by the pathologist. The subject identifier (i.e. the access number on slide labels) is coded and then the identifiers are removed after diagnostic tabulation. That is, the coded data is kept, hence, there is not any direct linked back to the subjects. These cases are reviewed by breast pathologist Dr. Nagi in collaboration with Shabnam Jaffer MD. at MSSM to reach a consensus.

This selection is made uniformly random, although preference is given to cases from the last 5 years. This allows access to more recent cases which are managed with modern clinical, radiological, surgical and pathological techniques. All patient populations, regardless of age, sex, or race, are included in the set. 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. However, histopathologically they are usually easily discernable 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. The images are taken using a 10 microscope objective lens with 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.

B. Results and Interpratation

We have calculated the accuracy of intensity-based approach, Delaunay-based approach, simple cell-graphs, hierarchical cell-graphs and hybrid-based approach and then compared them to each other in table III.

Intensity-based learning: In the intensity-based approach [37], [43], [44], [45] features are extracted from the gray-level or color histogram 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 in [36] using gray-level histograms, the sum and mean of the optical densities of the pixels located in a nucleus are computed and defined as the intensity-based features of the nucleus. Likewise we extracted intensity-based features by employing the RGB values of pixels in a tissue. For each color channel, we computed the mean, standard deviation, skewness and kurtosis of the pixel values of an image. The skewness of a channel is a measure of the asymmetry of the data around the sample mean and the kurtosis of each channel is a mesaure of how outlier-prone that channel is. Skewness is given by:

$$y = \frac{E(x-\mu)^k}{\delta^k}.$$

where μ is the mean of that channel, δ is the standard deviation, $E(t)$ is the expected value of the quantity t and $k = 3$. The kurtosis is given by the same formula with a k value of 4.

Delaunay triangulation: In order to quantify the spatial distribution of nuclei, Voronoi diagrams and their Delaunay triangulations are proposed in [46], [47] and [48]. 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 build the Delaunay triangulation on cell-clusters that we find in node identification step. Then we evaluate the metrics explained in section III for these graphs. These metrics are then given as the feature set to the classifier.

The choice of the parameters for graph generations affects the learning ratio significantly. For hierarchical graphs, we

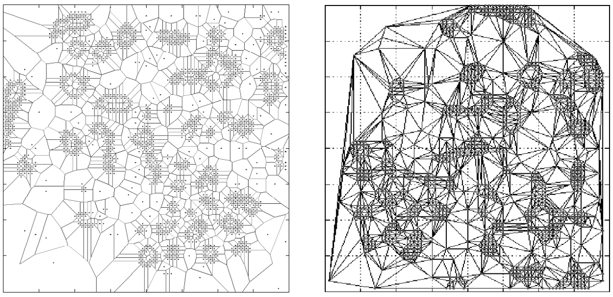


Fig. 3. (a)The Voronoi cells of the tissue. (b)The dual of the Voronoi diagram, Delaunay triangulation.

TABLE I

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

TABLE II

PROBABILISTIC CELL-GRAPHS

Link Threshold	5	6	7	8	9
Benign	92.0±3	88.7±4	89.2±4	91.6±2	91.1±3
InSitu	50.9±4	54.9±6	55.1±5	50.2±4	47.8±7
Invasive	79.2±4	75.9±4	74.6±7	77.1±4	78.1±3
Overall	74.5±2	73.2±3	72.6±4	73.1±1	72.9±2

TABLE III

COMPARISON OF THE TECHNIQUES

	Inten.	Delaun.	Prob.	Simple	Hier.	Hybrid
Benign	85.3	80.9	90.5	84.7	82.9	90.9
InSitu	50.9	16.3	51.8	51.6	75.6	57.3
Invasive	51.7	56.7	77.0	85.6	83.3	86.3
Overall	61.0	54.1	73.4	75.8	81.8	79.1

TABLE IV

DETAILED COMPARISIONS

	Intensity Prediction			Delaunay Prediction			Hierarchical Prediction		
Act	Ben	InS	Inv	Ben	InS	Inv	Ben	InS	Inv
Ben	85.3	10.3	4.4	80.9	10.3	8.8	82.9	7.3	9.8
InS	16.4	50.9	32.7	49.1	16.4	34.5	5.5	75.6	18.9
Inv	14.4	33.9	51.7	27.1	16.1	56.8	8.3	8.3	83.3

obtain the best result, 81.8%, 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. In table I 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, since the graphs start having cliques.

Table II shows the accuracy of the classifier with varying link values for probabilistic cell-graphs. We run our probabilistic cell-graph algorithm for 15 times to get a good estimate of the accuracy. Note that we do not need to run any of the other techniques more than one time since they are not probabilistic.

In table III we give the comparative results of the techniques discussed in the paper. Inten., Delaun, Prob. and Hier. are used as abbreviations for intensity, Delaunay, probabilistic cell-graphs, and hierarchical cell-graphs respectively.

From table III we see that intensity-based approach achieves a learning ratio of 61.0%. Delaunay triangulation of the cells produces worse results than the intensity-based approach even though it 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 average for link thresholds varying between 1 and 10. Probabilistic cell-graphs do not change the results significantly compared to the simple cell-graphs and 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 threshold and a fairly big grid size to find the clusters. For hierarchical graphs, after some point increasing the link threshold does not change the learning ratio as can be seen in I. This is because we obtain a complete graph where each node has a link to the other nodes. We have used a grid size of 10 which is able to capture the cell clusters. Using hierarchical graphs we obtained a learning ratio of 81.8%.

In hybrid-based approach we have 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%.

The over all learning ratio suggests that Hierarchical-graphs perform better than the other techniques presented in this paper. Besides, the learning ratio for in-situ case is 75.6% for hierarchical graphs and the next better result is 57.3%. This is because of the capability of hierarchical-graphs to capture lobular structures. In table IV we have also presented the confustion matrices of the techinques. In that table Ben, InS, Inv and Act are used as abbreviations for benign, in-situ, invasive, and actual (true) class. We see that hierarchical cell-graphs have false positive and false negative values smaller than 10%.

V. 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 in architecture. To capture this difference we introduce hierarchical graphs and obtaine the best learning ratio compared to the other techniques which is 81.8%.

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 on 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 hierarchical cell-graphs are much more accurate for breast tissues. However, we believe that accuracy can be improved further by increasing the data size and by improving the image segmentation.

REFERENCES

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

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

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

[4] R. Albert , H. Jeong, A.-L. Barabasi, ”Diameter of the World-Wide Web”, *Nature*, vol. 401, 1999, pp. 130-131.

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

[6] A.N. 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”, *IEEE Trans. Information Technology in Biomedicine*, vol. 2, no. 3, 1998, pp. 197-203.

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

[8] A-L. Barabasi, ”Linked: The New Science of Networks”, *Perseus Books Group; 1ST edition, 2002.*.

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

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

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

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

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

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

[15] M. Faloutsos, P. Faloutsos, and C. Faloutsos, ”On power-law relationships of the Internet topology”, *in Proceedings of ACM/SIGCOMM*, 1999, pp. 251-262.

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

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

[18] T.R. Golub, D.K. Slonim, 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. ”Molecular classification of cancer: class discovery and class prediction by gene expression monitoring”, *Science*, 286, 1999, pp. 531537.

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

[20] C. Gunduz, B. Yener, and S. H. Gultekin, ”The cell graphs of cancer”, *Bioinformatics*, vol. 20 2004, i145-i151.

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

[22] 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, 2001, pp. 233-239.

[23] H. Jeong, Tombor, R. Albert, Z. N. Oltvai, A.-L. Barabasi, ”The large-scale organization of metabolic networks”, *Nature* vol. 407, 2000, pp. 651-654.

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

[25] F. Liljeros, C. R. Edling, L. A. N. Amaral, H. E. Stanley, and Y. Aberg, ”The web of human sexual contacts”, *Nature*, vol.411 pp. 907-908.

[26] S. Milgram, ”The small-world problem”, *Psychology Today*, 1967, vol. 2 pp. 61-67.

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

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

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

[30] 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, 1997, pp. 68-75.

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

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

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

[34] S. Wasserman and K. Faust, ”Social network analysis: methods and applications”, *Cambridge University Press*, 1994.

[35] D. Watts and S. Strogatz, ”Collective dynamics of small-world networks”, *Nature* vol. 393, 1998, pp. 440-442.

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

[37] B. Weyn, G. Van de Wouwer, A. Van Daele, P. Scheunders, D. Van Dyck, E. Van Marck, and W. Jacob, ”Automated breast tumor diagnosis and grading based on wavelet chromatin texture description”, *Cytometry*, vol. 33, 1998, pp.32-40.

[38] W.H. Wolberg, W.N. Street, D.M. Heisey, and O.L. Mangasarian, ”Computer-Derived Nuclear Features Distinguish Malignant from Be-

nign Breast Cytology”, *Human Pathology*, vol. 26, no. 7, 1995, pp. 792-796.

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

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

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

[42] ”http:info.med.yale.edu/wmkeck/prochem/procmald.htm”

[43] M. Wiltgen, A. Gerger, and J. Smolle, ”Tissue counter analysis of benign common nevi and malignant melanoma”, *International Journal of Medical Informatics*, vol. 69, 2003, pp. 17-28.

[44] F. Schnorrenberg, C.S. Pattichis, C.N. Schizas, K. Kyriacou, and M. Vassiliou, ”Computer-aided classification of breast cancer nuclei”, *Technology and Health Care Journal*, vol. 4, 1996, pp.147-161.

[45] Z.H. Zhou, Y. Jiang, Y.B. Yang, and S.F. Chen, ”Lung cancer cell identification based on artificial neural network ensembles”, *Artificial Intelligence in Medicine, vol. 24, 2002, pp. 25-36.*

[46] S. J. Keenan, J. Diamond, W.G. McCluggage, H. Bharucha, D. Thompson, B.H. Bartels, P.W. Hamilton, ”An automated machine vision system for the histological grading of cervical intraepithelial neoplasia (CIN)”, Journal of Patholology, vol. 192, 2000, pp. 351-362.

[47] H.-K. Choi, T. Jarkrans, E. Bengtsson, J. Vasko, K. Wester, P.-U. Malmstrom, C. Busch, ”Image analysis based grading of bladder carcinoma. Comparison of object, texture and graph based methods and their reproducibility”, Analytical Cellular Pathology, vol. 15, 1997, pp. 1-18.

[48] B. Weyn, G. Van de Wouwer, S. Kumar-Singh, A. Van Daele, P. Scheunders, E. Van Marck, W. Jacob, ”Computer-assisted differential diagnosis of malignant mesothelioma based on syntactic structure analysis”, Cytometry, vol. 35, 1999, pp. 23-29.